Fast Iterative Solution of the Bethe-Salpeter Eigenvalue Problem using Low-Rank and QTT Tensor Approximation

Peter Benner* Sergey Dolgov** Venera Khoromskaia◇
Boris N. Khoromskij§

Abstract

In this paper, we propose and study two approaches to approximate the solution of the Bethe-Salpeter equation (BSE) by using structured iterative eigenvalue solvers. Both approaches are based on the reduced basis method and low-rank factorizations of the generating matrices. We also propose to represent the static screen interaction part in the BSE matrix by a small active sub-block, with a size balancing the storage for rank-structured representations of other matrix blocks. We demonstrate by various numerical tests that the combination of the diagonal plus low-rank plus reduced-block approximation exhibits higher precision with low numerical cost, providing as well a distinct two-sided error estimate for the smallest eigenvalues of the Bethe-Salpeter operator. The complexity is reduced to $\mathcal{O}(N_b^2)$ in the size of the atomic orbitals basis set, N_b, instead of the practically intractable $\mathcal{O}(N_b^6)$ scaling for the direct diagonalization. In the second approach, we apply the quantized-TT (QTT) tensor representation to both, the long eigenvectors and the column vectors in the rank-structured BSE matrix blocks, and combine this with the ALS-type iteration in block QTT format. The QTT-rank of the matrix entities possesses almost the same magnitude as the number of occupied orbitals in the molecular systems, $N_o < N_b$, hence the overall asymptotic complexity for solving the BSE problem by the QTT approximation is estimated by $\mathcal{O}(\log(N_o)N_o^2)$. We confirm numerically a considerable decrease in computational time for the presented iterative approaches applied to various compact and chain-type molecules, while supporting sufficient accuracy.

Key words: Bethe-Salpeter equation, Hartree-Fock calculus, tensor decompositions, quantized-TT format, model reduction, structured eigensolvers, low-rank matrix.
AMS Subject Classification: 65F30, 65F50, 65N35, 65F10

*Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstr. 1, D-39106 Magdeburg, Germany (benner@mpi-magdeburg.mpg.de)

**University of Bath, The Avenue, Bath, BA2 7AY, United Kingdom (S.Dolgov@bath.ac.uk).

◇Max Planck Institute for Mathematics in the Sciences, Leipzig; Max Planck Institute for Dynamics of Complex Systems, Sandtorstr. 1, D-39106 Magdeburg, Germany (vekh@mis.mpg.de).

§Max Planck Institute for Mathematics in the Sciences, Inselstr. 22-26, D-04103 Leipzig, Germany (bokh@mis.mpg.de).

1 Introduction

The Bethe-Salpeter equation (BSE) [46], [17] offers one of the commonly used mathematical models for *ab initio* computation of the absorption spectra for molecules or surfaces of solids, see also [51, 43, 37, 47, 29, 3, 42]. The BSE approach leads to the challenging computational task of the solution of a large eigenvalue problem for a fully populated (dense) matrix, that, in general, is non-symmetric. The size of the BSE matrix scales quadratically $\mathcal{O}(N_b^2)$ in the size N_b of the atomic orbitals basis sets, commonly used in *ab initio* electronic structure calculations. Hence, the direct diagonalization of $\mathcal{O}(N_b^6)$-complexity becomes prohibitive even for moderate size molecules.

Traditional methods for computer simulation of excitation energies for molecular systems require large computational facilities. Therefore there is a steady need for new algorithmic approaches for calculating the absorption spectra of molecules with less computational cost and having a good potential for application to larger systems. Recent tensor-structured methods for real-space electronic structure calculations provide cost-effective algorithms based on low-rank data sparse representations, which are transparent in implementation and suitable for MATLAB® on a laptop.

Conceptually, this paper continues the previous article [6], where a reduced basis approach to the Bethe-Salpeter algebraic eigenvalue problem (EVP) was introduced based on the idea of low-rank plus diagonal approximation to the BSE matrix blocks, which leads to a reduction of computational cost for a number of smallest in modulus eigenvalues from $\mathcal{O}(N^6)$ to $\mathcal{O}(N^2)$. The possibility for such an approximation of the BSE matrix blocks is suggested by the output of tensor-structured solvers for the Hartree-Fock (HF) eigenvalue problem [19, 21, 25, 22]. It provides not only the full set of eigenvalues and the coefficients for the expansion of molecular orbitals in a given basis for the ground state energy, but also an efficient representation of the two-electron integrals (TEI) tensor in a form of a low-rank Cholesky factorization[1] [23, 21].

Using the factorized TEI we applied and studied in [6] the approximate numerical solution of the BSE problem by a reduced basis method which included two steps. First, the diagonal plus low-rank approximation to blocks in the 2×2 BSE block matrix is calculated, enabling an easier partial eigenvalue solver for a large simplified system relying only on matrix-vector multiplications with rank-structured matrices. Second, the reduced basis approach was applied, via projection of the exact BSE matrix onto a *reduced basis*, constructed by the eigenvectors of the simplified eigenvalue problem. In our construction of the BSE matrix blocks, we use the particular description of the related quantities in the BSE matrix presented in [41], where the noninteracting Green's function was utilized.

In this paper, we propose and study two approaches to approximate the solution of the Bethe-Salpeter eigenvalue problem by using structured iterative solvers. Both are based on low-rank factorizations in the generating matrices [6].

First, we consider iterative schemes for computing several tens of the smallest in modulus eigenvalues for both the BSE problem and its Tamm-Dancoff approximation (TDA), based on the full representation of the eigenvectors and low-rank approximations of the BSE matrix

[1]The tensor-structured calculation of the TEI is designed by using a nonstandard "black-box" density fitting scheme and efficient 3D tensor-product convolution with the Newton kernel in 1D complexity. Fine 3D grids of the order of 10^{15} provide high accuracy, all algorithms are implemented in MATLAB on a laptop.

blocks. The most efficient subspace iteration is based on the application of the matrix inverse, which for our matrix formats can be evaluated in an efficient way using the Sherman-Morrison-Woodbury formula. As discussed in [6], the method reduces the numerical expense of the direct diagonalization down to $\mathcal{O}(N_b^2)$ in the size of the atomic orbitals basis set, N_b. The numerical experiments show that this method is economical up to small amino acids, where the numerical cost for computing several hundreds of eigenvalues decreases by orders of magnitude.

In the second approach, a reduction of the numerical cost in the case of large system size is achieved by adapting an ALS-type iteration (in particular, the DMRG iteration) for computing the eigenvectors in the block-QTT tensor representation [11]. The application of the QTT-approximation is motivated by the observation [21] that the generating Cholesky factors in the TEI tensor exhibit average QTT-ranks proportional only to the number of occupied orbitals in the molecular system, N_o, and independent of the total BSE matrix size, $\mathcal{O}(N_b^2)$. For eigenvectors in the block-QTT format, the QTT ranks are even smaller, typically proportional to the number of the sought eigenvectors, which makes this approach to solving the BSE eigenvalue problem very competitive. Contrarily to the conventional QTT matrix representations, in this paper we approximate only the columns in the Cholesky factor of a low-rank part of the BSE matrix in the QTT format, thus keeping the low-rank form $V = LL^T$ and low rank QTT structure for L simultaneously. This allows to avoid the prohibitive increase of the QTT matrix rank.

Instead of the problematic rank approximation to the statically screened interaction part of the BSE matrix, which complicated the trade-off between low-rank and accuracy requirements noticed in [6], here we propose to represent this part by a small active sub-block, with a size balancing the storage for rank-structured representations of other matrix blocks. We demonstrate that this combination of low-rank plus reduced-block approximation exhibits at least one order of magnitude higher precision at a similar low numerical cost[2]. Moreover, we observe a distinct two-sided error estimate for some tens of the smallest BSE eigenvalues, with the upper bounds resulting from the reduced basis problem and the lower bounds from the simplified BSE system matrix with diagonal plus low-rank and reduced block structure.

Notice that methods for solving partial eigenvalue problems for matrices with a special structure as in the BSE setting are conceptually related to the approaches for Hamiltonian matrices [4, 7, 30, 15, 9], particularly to those based on minimization principles [1, 2]. The special class of BSE-type equations leads to the so-called complex J-symmetric matrices, which have been intensively studied in [8, 34, 33, 35, 5] with a particular focus on the BSE problem [5]. Various structured eigensolvers tailored for electronic structure calculations are discussed in [44, 45, 10, 36, 32, 49].

The rest of the paper is organized as follows. In Section 2 we recall the reduced basis approach to the BSE problem introduced in [6], based on low-rank factorization of the BSE matrix blocks. Next, in Section 3 we describe the enhanced structural representation of the BSE system matrix by the reduced-block approximation to the statically screened interaction sub-matrix. The enhanced structured approximation improves the accuracy of the reduced basis method as justified by numerical simulations. Section 4 describes structured iterative

[2]As it was shown in [6], for the pure low-rank approach, with moderate ϵ-truncation of the rank parameters, the average error in the eigenvalues is of the order of 0.1 eV.

solvers for the central part of the spectrum in the simplified auxiliary problem. Section 5 discusses the benefits of the structured iterative solver based on the QTT tensor approximation of vectors and matrices in the framework of ALS-type subspace iterations in block-QTT format. In particular, we present and analyze numerically an algorithm for solving the BSE problem in $\mathcal{O}(\log(N_o)N_o^2)$ complexity, where $N_o \le CN_b$ ($C \approx 0.1$) denotes the number of occupied molecular orbitals.

Numerical tests (in MATLAB) confirm the considerable decrease in computational time while attaining sufficient accuracy. The conclusions summarize the main results and devise directions for future work.

2 The reduced basis approach to the BSE problem revisited

The construction of the BSE matrix includes computations of several auxiliary quantities [41, 6] represented in terms of the energy spectra ε_j, $j = 1, \dots, N_b$, and the rank-R_B two-electron integrals (TEI) matrix projected onto the Hartree-Fock molecular orbital basis,

$$V = [v'_{ia',jb'}] \quad a', b' \in \mathcal{I}_v := \{N_o + 1, \dots, N_b\}, \quad i, j \in \mathcal{I}_o := \{1, \dots, N_o\},$$

where $V' = [v'_{ia',jb'}]$ is a submatrix of the full TEI matrix, N_b is the number of GTO basis functions and N_o denotes the number of occupied orbitals (see [21, 6] for more details).

The 2×2-block matrix representation of the Bethe-Salpeter equation is given by the following eigenvalue problem determining the excitation energies ω_n:

$$\begin{pmatrix} A & B \\ B^* & A^* \end{pmatrix} \begin{pmatrix} \mathbf{x}_n \\ \mathbf{y}_n \end{pmatrix} = \omega_n \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix} \begin{pmatrix} \mathbf{x}_n \\ \mathbf{y}_n \end{pmatrix}, \tag{2.1}$$

where the matrix blocks (of size $N_{ov} \times N_{ov}$, with $N_{ov} = N_o(N_b - N_o)$) are defined by

$$A = \boldsymbol{\Delta\varepsilon} + V - \widehat{W}, \quad B = V - \widetilde{W}. \tag{2.2}$$

Here, the diagonal part is given by the "energy" matrix

$$\boldsymbol{\Delta\varepsilon} = [\Delta\varepsilon_{ia,jb}] \in \mathbb{R}^{N_{ov} \times N_{ov}} : \quad \Delta\varepsilon_{ia,jb} = (\varepsilon_{a'} - \varepsilon_i)\delta_{ij}\delta_{ab},$$

where ε_i and $\varepsilon_{a'}$ are the eigenvalues of the related Hartree-Fock equation. Here, in the left-hand side, we shift the a', b' indices to 1, introducing $a = a' - N_o$ and $b = b' - N_o$, i.e., $a, b = 1, \dots, N_v = N_b - N_o$. The double indices (i, a) and (j, b) can be seen as single long indices ia, jb (and vice versa) using the standard lexicographic grouping, e.g., $ia = i + (a-1)N_o = 1, \dots, N_{ov}$. The system (2.2) and all classical algebraic operations are considered w.r.t. the univariate indices ia, jb. However, the double index notation remains useful for describing fine structures, such as submatrices.

The "energy" matrix can be represented in the Kronecker product form

$$\boldsymbol{\Delta\varepsilon} = I_{N_o} \otimes \mathrm{diag}\{\varepsilon_{a'} : a' \in \mathcal{I}_v\} - \mathrm{diag}\{\varepsilon_i : i \in \mathcal{I}_o\} \otimes I_{N_v},$$

where I_{N_o} and I_{N_v} are the identity matrices of the respective sizes. The matrices $\widetilde{W}$ and $\widehat{W}$ are obtained by certain transformations of the matrix V.

In the present paper, the atomic orbitals are real-valued which imposes that the matrices A and B in (2.1) are also real-valued. Hence, in what follows, we use the notation A^T instead of A^* etc.

The matrices V and $\widetilde{W}$ are proven to have small ϵ-rank[3] (see Lemmas 2.1, 2.2 and 3.1 in [6]). In particular, there holds

$$V \approx L_V L_V^T, \quad L_V \in \mathbb{R}^{N_{ov} \times R_V}, \quad R_V \le R_B, \tag{2.3}$$

with the rank estimates $R_V = R_V(\epsilon) = \mathcal{O}(N_b|\log\epsilon|)$ and $\mathrm{rank}(\widetilde{W}) \le \mathrm{rank}(V)$. The arguments in [6] are based on an assumption concerning the separation properties of the TEI tensor in the Hartree-Fock calculations using Gaussian type orbitals [23, 21], i.e., that the ϵ-rank of the TEI tensor represented in the atomic orbital basis satisfies $R_B(\epsilon) = \mathcal{O}(N_b|\log\epsilon|)$. This basic assumption was verified numerically in [23, 21] for all molecular systems considered there. It was also demonstrated for the matrix V, see [6, Figure 1]. Moreover, a rank behavior like $R_B = \mathcal{O}(N_b)$ is conventionally used in the literature on electronic structure calculations although analytic proofs of this fact remain out of reach.

It was found in [6] that the matrix $\widehat{W}$ can be approximated by the low-rank substitute only up to the limited precision ϵ_0, so that a computationally inexpensive (but not accurate enough) approach to get rid of this limitation may be the rank approximation with the constraints $\mathrm{rank}(\widehat{W}) \le \mathrm{rank}(V)$.

Matrices in the form (2.1) are called J-symmetric (which equals Hamiltonian structure for real matrices), see [5] for implications on the algebraic properties of the BSE matrix. Solutions of equation (2.1) come in pairs: excitation energies ω_n with eigenvectors $(\mathbf{x}_n, \mathbf{y}_n)$, and de-excitation energies $-\omega_n$ with eigenvectors $(\mathbf{y}_n^*, \mathbf{x}_n^*)$. The spectral problem (2.1) can be rewritten in the equivalent form

$$F\begin{pmatrix}\mathbf{x}_n\\ \mathbf{y}_n\end{pmatrix} \equiv \begin{pmatrix}A & B\\ -B^T & -A^T\end{pmatrix}\begin{pmatrix}\mathbf{x}_n\\ \mathbf{y}_n\end{pmatrix} = \omega_n \begin{pmatrix}\mathbf{x}_n\\ \mathbf{y}_n\end{pmatrix}. \tag{2.4}$$

The dimension of the matrix in (2.1) is $2N_oN_v \times 2N_oN_v$, where N_o and N_v denote the numbers of occupied and virtual orbitals, respectively. In general, N_oN_v is asymptotically of the order $\mathcal{O}(N_b^2)$, i.e., the spectral problem (2.1) may become computationally expensive even for moderate size molecules, say for $N_b \approx 100$. Indeed, the direct eigenvalue solver for (2.1) (full diagonalization) appears to be infeasible due to $\mathcal{O}(N_b^6)$ complexity scaling.

The main idea of the *reduced basis approach* introduced in [6] can be described as follows. Instead of solving the partial eigenvalue problem for finding, say, m_0 eigenpairs satisfying equation (2.4), we first solve the slightly simplified auxiliary spectral problem with a modified matrix F_0. The approximation F_0 is obtained from F by using low-rank approximations of the matrices

$$\widehat{W} \mapsto \widehat{W}_r = L_W L_W^\top, \quad \text{and} \quad \widetilde{W} \mapsto \widetilde{W}_r = YZ^\top \tag{2.5}$$

in the matrix blocks A and B, respectively, i.e., A and B are replaced by

$$A \mapsto A_0 := \boldsymbol{\Delta\varepsilon} + V - \widehat{W}_r \quad \text{and} \quad B \mapsto B_0 := V - \widetilde{W}_r, \tag{2.6}$$

[3]Conventionally, we define the matrix ϵ-rank as the result of the truncated SVD w.r.t. the threshold $\epsilon > 0$. Throughout the paper, ϵ denotes the rank truncation parameter.

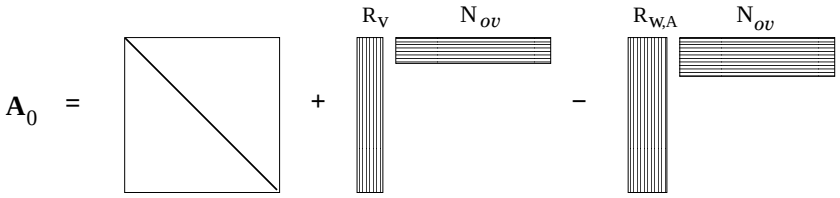

Figure 2.1: Diagonal plus low-rank structure of the matrix A_0.

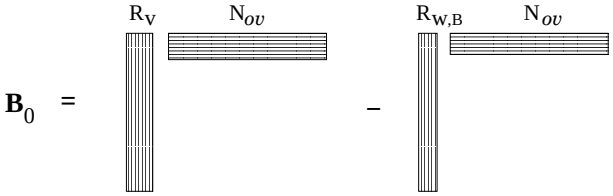

Figure 2.2: Low-rank structure of the matrix B_0.

where we assume for simplicity $\mathrm{rank}(\widehat{W}_r) \le r$ and $\mathrm{rank}(\widetilde{W}_r) \le r$. Here, we take into account that the matrix V, precomputed by the tensor-based Hartree-Fock solver [22], is already represented in the low-rank format (2.3) inherited from the Cholesky decomposition of the TEI matrix B, see [21, 6].

The simplified auxiliary problem reads

$$F_0 \begin{pmatrix} \mathbf{u}_n \\ \mathbf{v}_n \end{pmatrix} \equiv \begin{pmatrix} A_0 & B_0 \\ -B_0^T & -A_0^T \end{pmatrix} \begin{pmatrix} \mathbf{u}_n \\ \mathbf{v}_n \end{pmatrix} = \lambda_n \begin{pmatrix} \mathbf{u}_n \\ \mathbf{v}_n \end{pmatrix}. \tag{2.7}$$

This eigenvalue problem is a simplification of (2.4), since now the matrix blocks A_0 and B_0, defined in (2.6), are composed of diagonal and low-rank matrices, see Figures 2.1 and 2.2 illustrating the data sparse structure of these matrix blocks.

Having computed the set of eigenpairs $\{(\lambda_n, \psi_n) = (\lambda_n, (\mathbf{u}_n, \mathbf{v}_n)^T)\}$, corresponding to m_0 nearest to zero eigenvalues (middle part of the spectrum) of the modified problem (2.7), we solve the full eigenvalue problem for the reduced matrix (reduced model) obtained by projection of the initial equation onto the problem adapted small basis set $\{\psi_n\}_{n=1}^{m_0}$ of size m_0.

Now, define a matrix $G_0 = [\psi_1, \ldots, \psi_{m_0}] \in \mathbb{R}^{2N_{ov} \times m_0}$, whose columns are the eigenvectors of F_0, compute the related Galerkin and mass matrices by projection onto the reduced basis specified by the columns in G_0,

$$M_0 = G_0^T F G_0 \in \mathbb{R}^{m_0 \times m_0}, \quad S_0 = G_0^T G_0 \in \mathbb{R}^{m_0 \times m_0},$$

and then solve the reduced generalized eigenvalue problem of small size $m_0 \times m_0$,

$$M_0 \mathbf{q}_n = \gamma_n S_0 \mathbf{q}_n, \quad \mathbf{q}_n \in \mathbb{R}^{m_0}. \tag{2.8}$$

The portion of the m_0 eigenvalues γ_n, is expected to be very close to the lowest excitation energies ω_n $(n = 1, \ldots, m_0)$ in the initial spectral problem (2.1).

The so-called Tamm-Dancoff approximation (TDA) simplifies the equation (2.4) to a standard Hermitian eigenvalue problem

$$A \mathbf{x}_n = \mu_n \mathbf{x}_n, \quad \mathbf{x}_n \in \mathbb{R}^{N_{ov}}, \quad A \in \mathbb{R}^{N_{ov} \times N_{ov}} \tag{2.9}$$

with the factor two smaller matrix size N_{ov}. The reduced basis approach via low-rank approximation can be applied directly to the TDA equation, such that the simplified auxiliary problem reads

$$A_0\mathbf{u} = \lambda_n\mathbf{u},$$

where we are interested in finding the m_0 smallest eigenvalues.

Extensive numerical tests confirm the efficiency of the reduced model approach applied to both TDA and BSE problems for a number of single molecules, as well as to chain type systems [6].

Although the auxiliary eigenvalue equation (2.6) is much simpler than (2.4), the computation of dozens of eigenvectors from (2.7) corresponding to the middle part of the spectrum remains to be a challenging numerical task since the traditional algebraic solvers often converge slowly. As a remedy, one can perform matrix-vector operations with the inverse matrices A_0^{-1} or F_0^{-1}. The efficient construction and implementation of the structured matrix inverses A_0^{-1} and F_0^{-1} will be addressed in Section 4.

3 Approximating $\widehat{W}$ in reduced-block format

Taking into account limitations of the low-rank decomposition to the statically screened interaction matrix $\widehat{W}$, in what follows, we introduce an alternative way to the data-sparse approximation of this matrix based on its restriction to a smaller-size active sub-matrix.

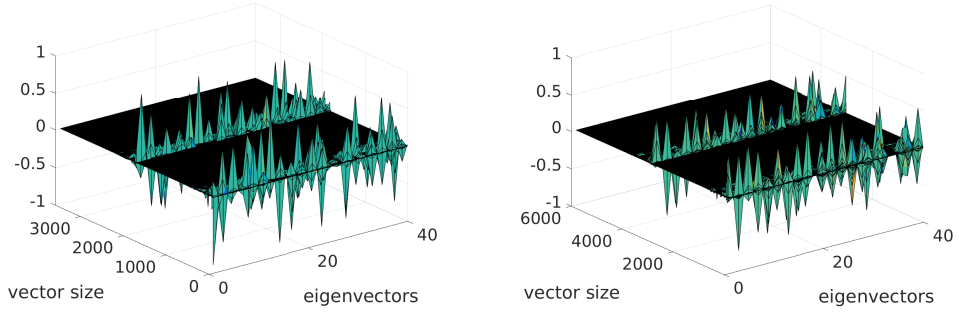

Figure 3.1: Visualizing the first m_0 BSE eigenvectors for the H_{32} chain (left) with $N_W = 554$, and Glycine amino acid molecule (right) with $N_W = 880$.

This approach is motivated by the numerical consideration (observed for all molecular systems considered so far) that eigenvectors corresponding to the central part of the spectrum have dominating components supported by a rather small part of the full index set of size $2\,N_{ov}$, see Figure 3.1 for $m_0 = 30$. Indeed, their effective support is compactly located at the first "active" indexes $\{1, \ldots, N_W\}$ and $\{N_{ov}+1, \ldots, N_{ov}+N_W\}$ in the respective blocks, where $N_W \ll N_{ov}$.

We define the selected sub-matrix $\widehat{W}_b$ in $\widehat{W}$, by keeping the balance between the storage size for the active sub-block $\widehat{W}_b$ and the storage for the matrix V. Since the storage and numerical complexity of the rank-R_V matrix V is bounded by $2\,R_V\,N_{ov}$, we control the size of the restricted $N_W \times N_W$ block $\widehat{W}_b$ by the relation

$$N_W = C_W\sqrt{2\;R_V\;N_{ov}}, \tag{3.1}$$

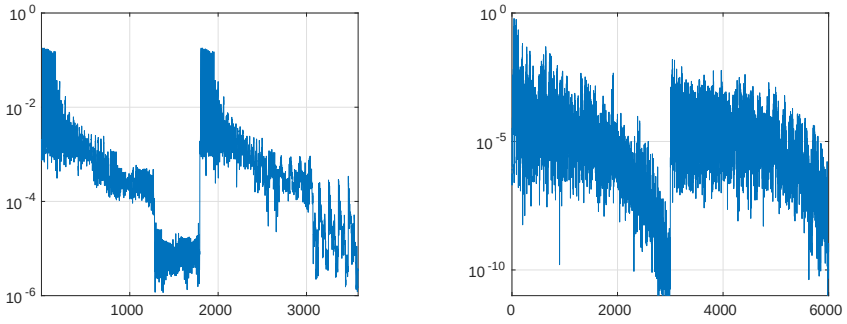

Figure 3.2: Visualizing the average decay of BSE eigenvectors at logarithmic scale corresponding to the m_0 smallest eigenvalues for the H_{32} chain (left) and the Glycine amino acid molecule (right).

where the constant C_W is close to 1. The approximation error introduced due to the corresponding matrix truncation can be controlled by the choice of the constant C_W.

Figure 3.2 shows the decay of BSE eigenvectors at logarithmic scale computed by averaging over m_0 eigenvectors (corresponding to the smallest eigenvalues) by

$$\mathbf{e}_{m_0}(:,1) = \frac{1}{2}\log\left(\sum_{\alpha=1}^{m_0} z(:,\alpha)^2\right) \in \mathbb{R}^{2N_{ov}}, \qquad \text{where } z = (x,y), \tag{3.2}$$

for the same molecular structures as in Figure 3.1. We notice that this figure confirms the computed choice of N_W for H_{32} chain, $N_W = 554$, and Glycine amino acid molecule, $N_W = 880$, with the truncation $\varepsilon = 0.1$ for low-rank approximation of other blocks.

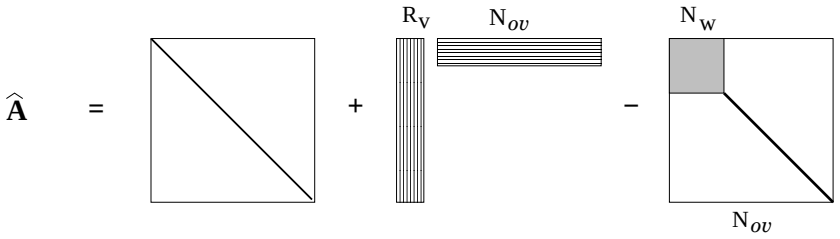

Figure 3.3: Diagonal plus low-rank plus reduced-block structure of the matrix $\widehat{A}$.

Keeping the diagonal in the matrix $\widehat{W}$ unchanged, we define the simplified matrix by $\widehat{W} \mapsto \widehat{W}_{N_W} \in \mathbb{R}^{N_{ov}\times N_{ov}}$, where

$$\widehat{W}_{N_W}(i,j) = \begin{cases} \widehat{W}(i,j), & i,j \le N_W \quad \text{or} \quad i=j, \quad \text{and} \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

The simplified matrix $\widehat{A}$ is then given by

$$A \mapsto \widehat{A} := \boldsymbol{\Delta\varepsilon} + V - \widehat{W}_{N_W}, \tag{3.4}$$

while the modified block B_0 remains the same as in (2.6). The corresponding structure of the simplified matrix $\widehat{A}$ is illustrated in Figure 3.3.

This construction guarantees that the storage and matrix-vector multiplication complexity for the simplified matrix block $\widehat{A}$ remains of the same order as that for the matrix V characterized by a low ε-rank. Table 3.1 demonstrates how the ratio N_W/N_{ov} evolves with the increasing problem size.

Molecule	H_2O	H_2O_2	N_2H_4	C_2H_5OH	H_{32}	$C_2H_5\ NO_2$	$C_3H_7\ NO_2$
N_W	114	269	266	460	552	880	1149
N_{ov}	180	531	657	1430	1792	3000	4488
N_W/N_{ov}	0.63	0.5	0.4	0.3	0.32	0.29	0.25

Table 3.1: The ratio N_W/N_{ov} for some molecules.

We modify the simplified matrix

$$F_0 \mapsto \widehat{F} \ \text{ by replacing } \ A_0 \mapsto \widehat{A} \text{ in } \ (2.7),$$

which leads to the corrections in the eigenvalues $\lambda_n \mapsto \widehat{\lambda}_n$ and eigenvectors $G_0 \mapsto \widehat{G} = [\psi_1, \ldots, \psi_{m_0}] \in \mathbb{R}^{2N_{ov} \times m_0}$ by solving the simplified problem,

$$\widehat{F}\psi_n = \widehat{\lambda}_n \psi_n,$$

defined by the low-rank plus block-diagonal approximation $\widehat{F}$ to the initial BSE matrix F. The corresponding eigenvalues $\widehat{\gamma}_n$ of the modified reduced system of the type (2.8), specified by the Galerkin and stiffness matrices

$$\widehat{M} = \widehat{G}^T F \widehat{G}, \quad \widehat{S} = \widehat{G}^T \widehat{G} \in \mathbb{R}^{m_0 \times m_0},$$

solve the eigenvalue problem of small size,

$$\widehat{M}\mathbf{q}_n = \widehat{\gamma}_n \widehat{S}\mathbf{q}_n, \quad \mathbf{q}_n \in \mathbb{R}^{m_0}, \tag{3.5}$$

by the direct diagonalization.

The following numerical examples illustrate the approximation error vs. the rank truncation parameter $\varepsilon > 0$ in the reduced basis method characterized by the choice of the constant C_W in the simplified matrix $\widehat{A}$ described in (3.4). Spectral data and errors are given in eV. Tables 3.2 (N_2H_4 molecule) and Table 3.3 (H_{16} chain) demonstrate the numerical errors $\widehat{\lambda}_1 - \omega_1$ and $\widehat{\gamma}_1 - \omega_1$ for the minimal BSE eigenvalue ω_1 for different rank truncation parameters ε, indicating the two-sided error estimates addressed in Remark 3.1 below.

Remark 3.1 *It is worth to note that numerical results indicate the important property observed for all molecular systems tested so far: the several close to zero eigenvalues $\widehat{\lambda}_k$ and $\widehat{\gamma}_k$ provide lower and upper bounds for the exact BSE eigenvalues ω_k, i.e.*

$$\widehat{\lambda}_k \le \omega_k \le \widehat{\gamma}_k, \quad k = 1, 2, \ldots, \underline{m}_0 \le m_0.$$

$C_W \setminus \epsilon$	0.2	0.1	0.05	0.01
0.8	-0.09; 0.006 (148)	-0.03; 0.04 (213)	-0.008; 0.014 (284)	-0.005; 0.0025 (406)
1.0	-0.1; 0.05 (185)	-0.036; 0.03 (266)	-0.015; 0.0076 (355)	-0.008; 0.0003 (507)
1.2	-0.1; 0.05 (222)	-0.04; 0.02 (320)	-0.017; 0.0038 (426)	$N_W = N_{ov}$

Table 3.2: N_2H_4, $N_{ov} = 657$. Errors $\widehat{\lambda}_1 - \omega_1$; $\widehat{\gamma}_1 - \omega_1$ (in eV), vs. ε and C_W. Here N_W for corresponding C_W and ϵ is given in brackets.

$C_W \setminus \varepsilon$	0.2	0.1	0.05	0.01
0.8	-0.23; 0.13 (131)	-0.054; 0.08 (157)	-0.047; 0.06 (168)	-0.006; 0.02 (200)
1.0	-0.28; 0.06 (164)	-0.1; 0.01 (196)	-0.073; 0.015 (210	-0.005; 0.02 (250)
1.2	-0.31; 0.01 (197)	-0.1; 0.01 (236)	-0.074; 0.013 (251)	-0.001; 0.005 (301)

Table 3.3: H_{16} chain, $N_{ov} = 448$: Errors $\widehat{\lambda}_1 - \omega_1$; $\widehat{\gamma}_1 - \omega_1$ (in eV), vs. ε and C_W; N_W is given in brackets.

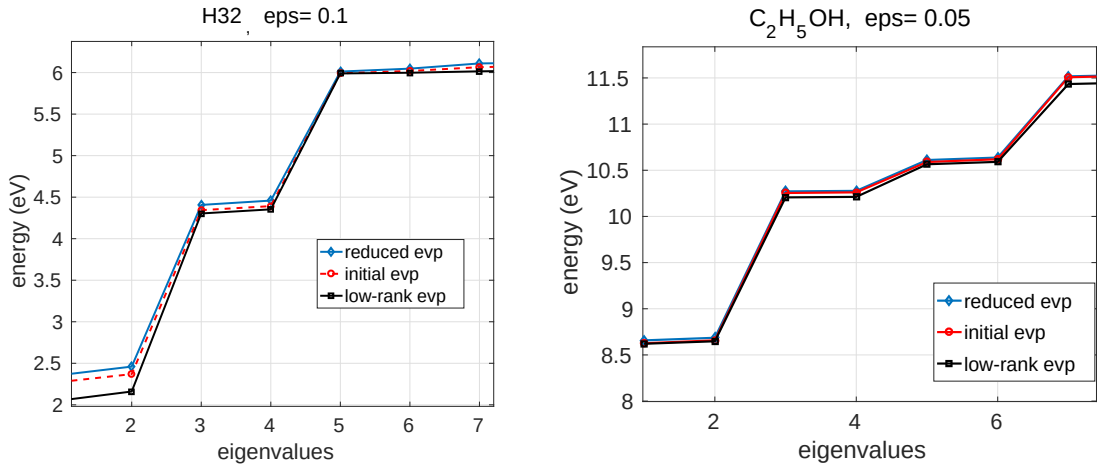

Figure 3.4: Two-sided bounds for the BSE excitation energies for the H_{32} chain (left) and $C_2H_5NO_2$ molecule (right).

The upper bound via the eigenvalues $\widehat{\gamma}_k$ can be explained by the variational form of the reduced problem setting. However, the understanding of the lower bound property, when using the output from the simplified system, addresses an interesting open problem.

Figure 3.4 demonstrates the two-sided error estimates declared in Remark 3.1. Here the "black" line represents the eigenvalues for the auxiliary problem of the type (2.7), but with the modified matrix $\widehat{F}$, while the blue line represents the eigenvalues of the reduced equation (3.5) of the type (2.8) with the Galerkin matrices $\widehat{M}$ and $\widehat{S}$.

Figures 3.5 and 3.6 represents examples of upper and lower bounds for the whole sets of $m_0 \leq 250$ eigenvalues for larger molecules. We observe that the lower bound is violated only for few larger excitation energies at the level below the truncation error ϵ.

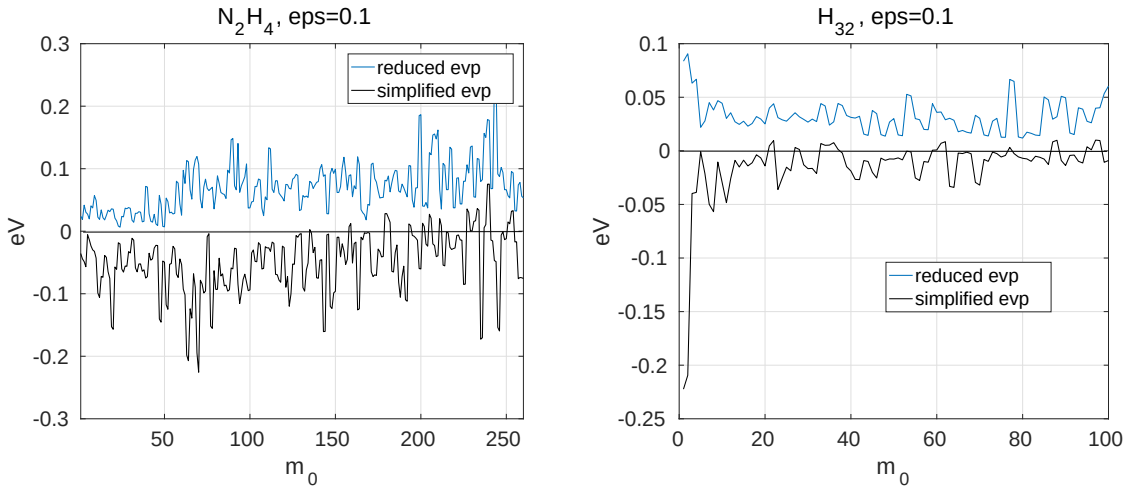

Figure 3.5: The errors (in eV) in eigenvalues for simplified and reduced schemes: the N_2H_4 for $m_0 = 260$ eigenvalues (left), the H_{32} chain (right) with $m_0 = 100$. Zero level designates the solution of the initial BSE problem.

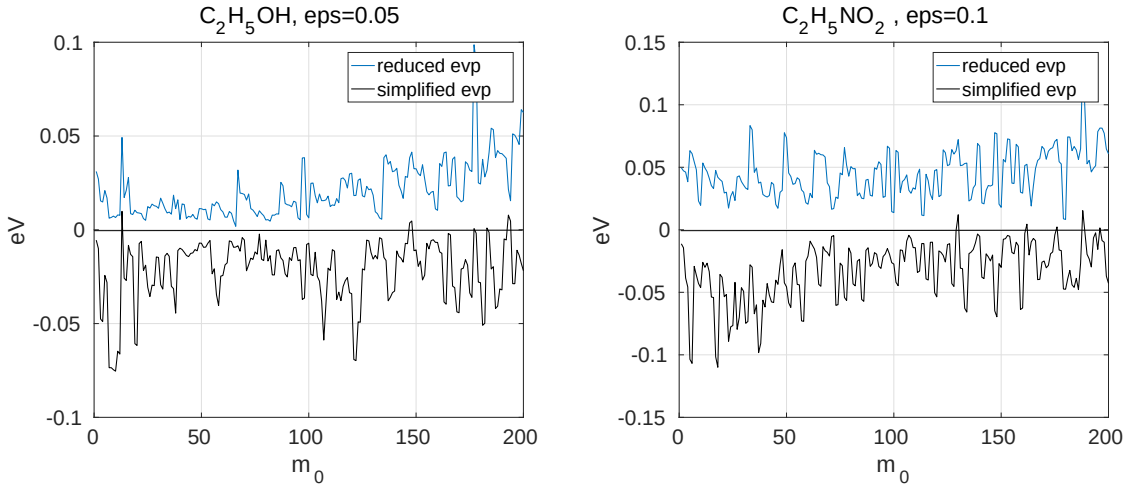

Figure 3.6: The errors (in eV) of the m_0 smallest eigenvalues for simplified and reduced schemes: Ethanol C_2H_5OH (left), Glycine amino acid $C_2H_5NO_2$ (right).

We conclude that the reduced basis approach based on the modified auxiliary matrix $\widehat{M}$ via reduced-block Ansätze (3.4), provides considerably better accuracies $\omega_n - \widehat{\gamma}_n$ than that for γ_n corresponding to matrix M_0. Table 3.4 compares the accuracies for first eigenvalues of the reduced BSE problem based on the "pure" low-rank approximation $|\omega_1 - \gamma_1|$ from equation (2.8) with those resulting from combined block plus low-rank approximation $|\omega_1 - \widehat{\gamma}_1|$, computed for several molecules.

Molecule	H_2O	N_2H_4	C_2H_5OH	$C_2H_5\ NO_2$	$C_3H_7\ NO_2$
BSE size	360^2	1314^2	2660^2	6000^2	8976^2
$\lvert\omega_1 - \gamma_1\rvert$	0.2	0.27	0.4	0.38	0.53
$\lvert\omega_1 - \widehat{\gamma}_1\rvert$	0.02	0.03	0.08	0.05	0.1

Table 3.4: Accuracies (in eV) for the reduced BSE problem eigenvalues for low-rank [6] $|\omega_1 - \gamma_1|$ and for block plus low-rank approximation $|\omega_1 - \widehat{\gamma}_1|$ to BSE matrices with the $\epsilon = 0.1$.

4 Iterative solver for central part of the spectrum

In this section we discuss the construction of an iterative solver for the partial eigenvalue problem in (2.7) focusing on rank-structured approximation of the matrix inverses A_0^{-1} and F_0^{-1}, further optimization of the sparsity pattern in $\widehat{W}$ and on the choice of the initial guess by using solutions of the TDA model.

4.1 Inverse iteration for diagonal plus low-rank matrix

Iterative eigenvalue solvers, such as Lanczos or Jacobi-Davidson methods, are quite efficient in approximation of the largest eigenvalues, but may suffer from slow convergence if applied for computation of the smallest or intermediate eigenvalues. We are interested in both of these scenarios. There are both positive and negative eigenvalues in (2.7), and we need the few ones with the smallest magnitude. In the TDA model (2.9), we solve a symmetric positive definite problem $A_0\mathbf{u} = \lambda_n\mathbf{u}$, but again the smallest eigenvalues are required.

In both cases, the remedy is to invert the system matrix, so that the eigenvalues of interest become largest. The MATLAB interface to ARPACK (procedure `eigs`) [31] assumes by default that the user-defined function solves a linear system with the matrix instead of multiplying it, when the smallest eigenvalues are requested. In our case, we can implement this efficiently, since the matrix consists of an easily invertible part (diagonal), plus a low-rank correction, and hence we can use the Sherman-Morrison formula [50].

To shorten the notation, we set up the rank-r decompositions following (2.5), $\widehat{W}_r = L_W L_W^T$, $\widetilde{W}_r = YZ^T$, and define

$$\begin{aligned} A_0 &= \mathbf{\Delta\varepsilon} + PQ^T, \qquad P = \begin{bmatrix} L_V & L_W \end{bmatrix}, \quad Q = \begin{bmatrix} L_V & -L_W \end{bmatrix}, \\ B_0 &= \Phi\Psi^T, \qquad\qquad \Phi = \begin{bmatrix} L_V & Y \end{bmatrix}, \quad \Psi = \begin{bmatrix} L_V & -Z \end{bmatrix}. \end{aligned} \tag{4.1}$$

taking into account (2.3).

Consider first the TDA model (2.9). The Sherman-Morrison formula for A_0 in (4.1) reads

$$A_0^{-1} = \boldsymbol{\Delta\varepsilon}^{-1} - \boldsymbol{\Delta\varepsilon}^{-1} P \left(I + Q^T \boldsymbol{\Delta\varepsilon}^{-1} P\right)^{-1} Q^T \boldsymbol{\Delta\varepsilon}^{-1}. \tag{4.2}$$

Here the inner $2r \times 2r$ matrix $K = \left(I + Q^T \boldsymbol{\Delta\varepsilon}^{-1} P\right)^{-1}$ is small and can be computed explicitly at the expense $\mathcal{O}(r^3 + r^2 N_{ov})$. Hence, the matrix-vector product $A_0^{-1}\mathbf{u}_n$ requires multiplication by the diagonal matrix $\boldsymbol{\Delta\varepsilon}^{-1}$ and the low-rank matrix in the second summand. This amounts to the overall cost $\mathcal{O}(N_{ov} r)$.

To invert F_0, we first derive its LU decomposition. One can verify that

$$F_0 = \begin{bmatrix} A_0 & B_0 \\ -B_0^T & -A_0^T \end{bmatrix} = \begin{bmatrix} A_0 & 0 \\ -B_0^T & I \end{bmatrix} \begin{bmatrix} I & A_0^{-1} B_0 \\ 0 & S \end{bmatrix}, \qquad S = -A_0^T + B_0^T A_0^{-1} B_0. \tag{4.3}$$

To solve a system $F_0 \begin{bmatrix} \mathbf{z} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}$, we need one action of A_0^{-1} and of the inverse of the Schur complement S^{-1}. Indeed,

$$\begin{aligned} \tilde{\mathbf{z}} &= A_0^{-1}\mathbf{u}, \quad \tilde{\mathbf{y}} = \mathbf{v} + B_0^T \tilde{\mathbf{z}}, \\ \mathbf{y} &= S^{-1}\tilde{\mathbf{y}}, \quad \mathbf{z} = \tilde{\mathbf{z}} - A_0^{-1} B_0 \mathbf{y}. \end{aligned} \tag{4.4}$$

Note that $A_0^{-1} B_0$ is a low-rank matrix and can be precomputed in advance. The action of A_0^{-1} is given by (4.2), so we address now the inversion of the Schur complement.

Plugging (4.2) into S, we obtain

$$S = -\boldsymbol{\Delta\varepsilon} - QP^T + \Psi\Phi^T A_0^{-1} \Phi\Psi^T = -(\boldsymbol{\Delta\varepsilon} + Q_S P_S^T),$$

where

$$Q_S = \begin{bmatrix} Q & \Psi \left(\Phi^T \boldsymbol{\Delta\varepsilon}^{-1} P K Q^T \boldsymbol{\Delta\varepsilon}^{-1} \Phi - \Phi^T \boldsymbol{\Delta\varepsilon}^{-1} \Phi\right) \end{bmatrix}, \quad P_S = \begin{bmatrix} P & \Psi \end{bmatrix}. \tag{4.5}$$

Therefore,

$$S^{-1} = -\left(\boldsymbol{\Delta\varepsilon}^{-1} - \boldsymbol{\Delta\varepsilon}^{-1} Q_S K_S P_S^T \boldsymbol{\Delta\varepsilon}^{-1}\right), \quad K_S = \left(I + P_S^T \boldsymbol{\Delta\varepsilon}^{-1} Q_S\right)^{-1}. \tag{4.6}$$

Keeping intermediate results in these calculations, we can trade off the memory against the CPU time. The computational cost of (4.5) and then (4.6) is again bounded by $\mathcal{O}(r^2 N_{ov})$, while the implementation of (4.4) takes $\mathcal{O}(r N_{ov})$ operations.

We have thus proven the following statement.

Lemma 4.1 (Complexity of the diagonal plus low-rank approach) *Let the rank parameters in the decompositions of V, $\widehat{W}$ and $\widetilde{W}$ not exceed r. Then the rank structured representations of the inverse matrices A_0^{-1} and F_0^{-1} can be precomputed with the overall cost $\mathcal{O}(N_{ov} r^2)$. The complexity for each inversion $A_0^{-1}\mathbf{u}$ or $F_0^{-1}\mathbf{w}$ is bounded by $\mathcal{O}(N_{ov} r)$.*

Lemma 4.1 indicates that for both, the BSE and TDA models, the asymptotic complexity for one iterative step is of the same order. Precomputation of intermediate matrices is described in Algorithm 1, and their use in the structured matrix inversion is shown in Algorithm 2 below.

Algorithm 1 Precomputation of parts of A_0^{-1} and F_0^{-1}

Require: $\mathbf{\Delta\varepsilon}$ and low-rank factors of V, $\widehat{W}_r$, $\widetilde{W}_r$ (2.5).

1: Assemble $P = \begin{bmatrix} L_V & L_W \end{bmatrix}$, $Q = \begin{bmatrix} L_V & -L_W \end{bmatrix}$, $\Phi = \begin{bmatrix} L_V & Y \end{bmatrix}$, $\Psi = \begin{bmatrix} L_V & -Z \end{bmatrix}$.
2: Compute $P_\varepsilon = \mathbf{\Delta\varepsilon}^{-1} P$, $Q_\varepsilon = \mathbf{\Delta\varepsilon}^{-1} Q$.
3: Compute $K = (I + Q^T P_\varepsilon)^{-1} \in \mathbb{R}^{2r \times 2r}$.
4: Compute $P_{\varepsilon K} = P_\varepsilon K$.
{Stop here if only A_0^{-1} is of interest}
5: Compute $\Phi_\varepsilon = \mathbf{\Delta\varepsilon}^{-1}\Phi$, $\Psi_\varepsilon = \mathbf{\Delta\varepsilon}^{-1}\Psi$.
6: Parts of Q_S: $\Phi_{\varepsilon P} = \Phi_\varepsilon^T P$, $\Phi_{\varepsilon Q} = Q^T \Phi_\varepsilon$.
7: Assemble $Q_{S\varepsilon} = \begin{bmatrix} Q_\varepsilon & \Psi_\varepsilon \left(\Phi_{\varepsilon P} K \Phi_{\varepsilon Q} - \Phi^T \Phi_\varepsilon\right) \end{bmatrix}$, $P_{S\varepsilon} = \begin{bmatrix} P_\varepsilon & \Psi_\varepsilon \end{bmatrix}$.
8: Compute $K_S = (I + \begin{bmatrix} P & \Psi \end{bmatrix}^T Q_{S\varepsilon})^{-1} \in \mathbb{R}^{4r \times 4r}$
9: Compute $Q_{S\varepsilon K} = Q_{S\varepsilon} K_S$. {For the Schur complement}
10: Compute $\Phi_{AB} = \mathbf{\Delta\varepsilon}^{-1}\Phi - P_{\varepsilon K}\left(Q_\varepsilon^T \Phi\right)$. {For $A_0^{-1} B_0$}

Algorithm 2 Solution of linear systems with A_0 and F_0

Require: Precomputed matrices $P_{\varepsilon K}, Q_\varepsilon, Q_{S\varepsilon K}, P_{S\varepsilon}, \Phi_{AB}$ from Alg. 1 and $\mathbf{\Delta\varepsilon}, \Phi, \Psi$.

Ensure: $\tilde{\mathbf{z}} = A_0^{-1}\mathbf{u}$ and $\begin{bmatrix} \mathbf{z} \\ \mathbf{y} \end{bmatrix} = F_0^{-1} \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}$

1: Apply the TDA inverse as $\tilde{\mathbf{z}} \equiv A_0^{-1}\mathbf{u} = \mathbf{\Delta\varepsilon}^{-1}\mathbf{u} - P_{\varepsilon K}\left(Q_\varepsilon^T \mathbf{u}\right)$.
{Stop here if only A_0^{-1} is of interest}
2: Compute $\tilde{\mathbf{y}} = \mathbf{v} + \Psi\left(\Phi^T \tilde{\mathbf{z}}\right)$ using (4.4)
3: Apply the Schur complement $\mathbf{y} \equiv S^{-1}\tilde{\mathbf{y}} = -\mathbf{\Delta\varepsilon}^{-1}\tilde{\mathbf{y}} + Q_{S\varepsilon K}\left(P_{S\varepsilon}^T \tilde{\mathbf{y}}\right)$.
4: Compute $\mathbf{z} = \tilde{\mathbf{z}} - \Phi_{AB}\left(\Psi^T \mathbf{y}\right)$.

Table 4.1 compares CPU times (sec) for full `eig` and the rank-structured iteration for TDA problem (2.9) in Matlab implementation. The rank-truncation threshold is $\varepsilon = 0.1$, the number of computed eigenvalues is $m_0 = 30$. The bottom line shows the CPU times (sec) of the `eigs` procedure applied with the inverse matrix-vector product $A_0^{-1}\mathbf{u}$ using Algorithm 2 (marked by "inv"). The other lines show results of the corresponding algorithms which used the traditional product $A_0\mathbf{u}$ (A_0 in the low-rank form). Notice that the results for Matlab version of LOBPCG by [27] are presented for comparison. We see that the inverse-based method is superior in all tests.

Remark 4.2 *Notice that the initial guess for the subspace iteration applied to the full BSE can be constructed, replicating the eigenvectors computed in the TDA model. It provides rather accurate approximation to the exact eigenvectors for the initial BSE system (2.4). In [6] it was shown numerically that the TDA approximation error $|\mu_n - \omega_n|$ of the order of 10^{-2} eV is achieved for the compact and extended molecules presented in Table 4.1.*

Table 4.2 compares CPU times (sec) for the full `eig`-solver and the rank-structured `eigs`-iteration applied to the inverse of simplified rank-structured BSE system (2.7).

Molecular syst.	H_2O	N_2H_4	C_2H_5OH	H_{32}	$C_2H_5\ NO_2$	H_{48}	$C_3H_7\ NO_2$
TDA size	180^2	657^2	1430^2	1792^2	3000^2	4032^2	4488^2
`eig`(A_0)	0.02	0.5	4.3	9.8	37.6	91	127.4
`lobpcg`(A_0)	0.22	0.6	5.4	2.77	18.2	5.6	34.2
`eigs`(A_0)	0.07	0.29	1.7	0.49	–	–	–
`eigs`(inv(A_0))	0.05	0.08	0.17	0.11	0.32	0.34	0.5

Table 4.1: Times (s) for eigenvalue problem solvers applied to TDA matrix ("–" means that the respective iteration did not converge).

Molecule	H_2O	N_2H_4	C_2H_5OH	H_{32}	$C_2H_5\ NO_2$	H_{48}	$C_3H_7\ NO_2$
N_o, N_b	5, 41	9 , 82	13, 123	16, 128	20, 170	24, 192	24, 211
BSE matrix size	360^2	1314^2	2860^2	3584^2	6000^2	8064^2	8976^2
`eig`(F_0)	0.08	4.2	33.7	68.1	274	649	903
`eigs`(inv(F_0))	0.13	0.28	0.7	0.77	2.2	2.3	3.9

Table 4.2: Times (s) for the simplified rank-structured BSE matrix F_0.

4.2 Inversion of the block-sparse matrices

If $\widehat{W}_{N_W}$ is kept in the block-diagonal form as in (3.4), inverting $\widehat{A} = \mathbf{\Delta\varepsilon} + V - \widehat{W}_{N_W}$ is also easy, similarly to the case (2.6). We can use the same Sherman-Morrison-Woodbury scheme as in Algorithms 1 and 2. To that end, we aggregate $\mathbf{\Delta\varepsilon}_W = \mathbf{\Delta\varepsilon} - \widehat{W}_{N_W}$, while in the low-rank factors, only $P = Q = L_V$ remains. After that, all calculations in Algorithms 1 and 2 are repeated unchanged, replacing all $\mathbf{\Delta\varepsilon}$ by $\mathbf{\Delta\varepsilon}_W$, where the latter is now a block-diagonal matrix.

The particular modifications for the enhanced algorithm are as follows. Let us split $\mathbf{\Delta\varepsilon} = \mathrm{blockdiag}(\mathbf{\Delta\varepsilon}_1, \mathbf{\Delta\varepsilon}_2)$, where $\mathbf{\Delta\varepsilon}_1$ has the size N_W, and $\mathbf{\Delta\varepsilon}_2 \in \mathbb{R}^{N'_W \times N'_W}$ with $N'_W = N_{ov} - N_W$ representing the remaining values. The same applies to $\widehat{W}_{N_W} = \mathrm{blockdiag}(W_b, \mathrm{diag}(w_2))$, where w_2 contains the elements on the diagonal of $\widehat{W}$ which do not belong to W_b. Then the implementation of the matrix inverse

$$\mathbf{\Delta\varepsilon}_W^{-1} = \mathrm{blockdiag}((\mathbf{\Delta\varepsilon}_1 - W_b)^{-1}, (\mathbf{\Delta\varepsilon}_2 - \mathrm{diag}(w_2))^{-1}) \tag{4.7}$$

requires inversion of an $N_W \times N_W$ dense matrix, and a diagonal matrix of size $N'_W = N_{ov} - N_W$. Since N_W is chosen small, the complexity of this operation is moderate. Now all steps requiring multiplication with $\mathbf{\Delta\varepsilon}^{-1}$ in Algorithms 1–2 can be substituted by (4.7). The numerical complexity of the new inversion scheme is estimated in the next lemma.

Lemma 4.3 (Complexity of the reduced-block algorithm) *Suppose that the rank parameters in the decomposition of V and $\widetilde{W}$ do not exceed r and the block-size N_W is chosen from the equation (3.1). Then the rank structured plus reduced-block representations of the inverse matrices $\widehat{A}^{-1}$ and $\widehat{F}^{-1}$ can be set up with the overall cost $\mathcal{O}(N_{ov}^{3/2} r^{3/2} + N_{ov} r^2)$. The complexity of each inversion $\widehat{A}^{-1}\mathbf{u}$ or $\widehat{F}^{-1}\mathbf{w}$ is bounded by $\mathcal{O}(N_{ov} r)$.*

Molecular syst.	H_2O	N_2H_4	C_2H_5OH	H_{32}	$C_2H_5\ NO_2$	H_{48}	$C_3H_7\ NO_2$
TDA size	180^2	657^2	1430^2	1792^2	3000^2	4032^2	4488^2
`eigs`$(\mathrm{inv}(\widehat{A}))$	0.07	0.09	0.25	0.77	0.54	3.0	1.0
`eigs`$(\mathrm{inv}(\widehat{F}))$	0.21	0.37	1.11	1.10	2.4	2.92	4.6
BSE vs. $\widehat{F}$: $\lvert\widehat{\gamma}_1 - \omega_1\rvert$	0.02	0.03	0.08	0.07	0.05	0.10	0.1

Table 4.3: Block-sparse matrices: times (s) for eigensolvers applied to TDA and BSE systems. Bottom line shows the error (eV) for the case of block-sparse approximation to the diagonal matrix block $\widehat{A}$, $\varepsilon = 0.1$.

Proof. Inversion of the $N_W \times N_W$ dense block in (4.7) requires $\mathcal{O}(N_W^3)$ operations. Hence, the condition (3.1) ensures that the cost of setting up the matrix (4.7) is bounded by $\mathcal{O}(N_{ov}^{3/2} r^{3/2})$. After that, multiplication of (4.7) by an $N_{ov} \times r$ matrix (e.g. in Line 2 of Algorithm 1) requires $\mathcal{O}(N_W^2 r + N_W' r) = \mathcal{O}(N_{ov}(r^2 + r))$ operations. In Algorithm 2, multiplication of (4.7) by a vector is performed with $\mathcal{O}(N_W^2 + N_W') = \mathcal{O}(N_{ov} r)$ cost. The complexity of the other steps is the same as in Lemma 4.1. ∎

Numerical illustrations for the enhanced data sparsity are presented in Table 4.3.

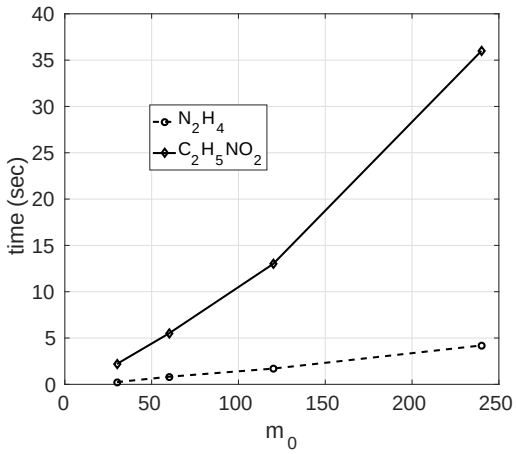

Figure 4.1: CPU times vs. m_0 for N_2H_4 (dashed line) and $C_2H_5NO_2$ (solid line) and C_2H_5OH (doted line) molecules.

Notice that the performance of the low-rank and block-sparse solvers is comparable, but the second one provides better sparsity and higher accuracy in the computed eigenvalues, see §3. It is remarkable that the approach, based on the inverse iteration applied to the diagonal plus low-rank plus reduced-block approximation, outperforms the full eigenvalue solver by several orders of magnitude, see Tables 4.2 and 4.3.

The data in previous tables corresponds to the choice $m_0 = 30$. Figure 4.1 indicates a merely linear increase in the computational time with respect to the increasing value of m_0.

5 Solving BSE spectral problems in the QTT format

In the recent years, the tensor methods were recognized as the powerful techniques that allows to enhance the traditional numerical methods by using low-rank separable representations of discretized functions and operators. In this section we introduce the main notations and definitions of the rank-structured tensor formats which will be used for data-sparse representation of large matrices and long vectors arising from the BSE problem. The approach is based on the so-called quantized-TT (QTT) low-rank tensor approximation [24] of vectors and matrices described in §5.2.

5.1 Rank-structured representation of multi-dimensional tensors

A real tensor of order d is defined as an element of the finite dimensional Hilbert space $\mathbb{W}_{\mathbf{m}} = \mathbb{R}^{M_1 \times ... \times M_d}$ composed of the d-fold, $M_1 \times ... \times M_d$ real-valued arrays (tensors), where $\mathbf{m} = (M_1, \ldots, M_d)$, and $I_\ell := \{1, \ldots, M_\ell\}$, $\ell = 1, \ldots, d$. A tensor $\mathbf{A} \in \mathbb{R}^{M_1 \times ... \times M_d}$ is represented entry-wise by

$$\mathbf{A} = [a(i_1, ..., i_d)] \equiv [a(\mathbf{i})] \equiv [a_{i_1,...,i_d}] \equiv [a_{\mathbf{i}}] \quad \text{with} \quad \mathbf{i} \in \mathcal{I} = I_1 \times ... \times I_d.$$

The Euclidean scalar product, $\langle \cdot, \cdot \rangle : \mathbb{W}_{\mathbf{m}} \times \mathbb{W}_{\mathbf{m}} \to \mathbb{R}$, is defined by

$$\langle \mathbf{A}, \mathbf{B} \rangle := \sum_{\mathbf{i} \in \mathcal{I}} a_{\mathbf{i}} b_{\mathbf{i}}, \quad \mathbf{A}, \mathbf{B} \in \mathbb{W}_{\mathbf{m}}.$$

The storage size for a dth order tensor scales exponentially in d, $\dim(\mathbb{W}_{\mathbf{m}}) = M_1 \cdots M_d$, that causes the so-called "curse of dimensionality". In this section, for ease of presentation we assume $M_\ell = M$ for $\ell = 1, \ldots, d$.

The efficient low-parametric representations of dth order tensors can be realized by using low-rank separable decompositions (formats). The commonly used canonical and Tucker tensor formats [28] are constructed by linear combination of the simplest separable elements given by rank-1 tensors,

$$\mathbf{U} = \mathbf{u}^{(1)} \otimes \cdots \otimes \mathbf{u}^{(d)} \in \mathbb{R}^{M_1 \times \cdots \times M_d}, \quad \mathbf{u}^{(\ell)} \in \mathbb{R}^{M_\ell},$$

with entries $u_{i_1,...i_d} = u^{(1)}_{i_1} \cdots u^{(d)}_{i_d}$, which can be stored using dM numbers.

Tensor-structured numerical methods for PDEs were particularly initiated by employment of the canonical and Tucker tensor formats in grid based "ab initio" electronic structure calculations, namely, for accurate evaluation of the 3D convolution integrals with the Newton kernel, see [22] and references therein. The literature overview on multi-linear algebra and tensor numerical methods for PDEs can be found, for example, in [28, 26, 16, 12, 22].

In this paper we apply the factorized representation of dth order tensors in the tensor train (TT) format [40], which is a particular case of the *matrix product states* (MPS) decomposition [53, 52, 48]. The latter was introduced long since in the physics community and successfully applied in quantum chemistry computations and in spin systems modeling. For a given rank parameter $\mathbf{r} = (r_1, \ldots, r_{d-1})$, and the respective index sets $J_\ell = \{1, ..., r_\ell\}$ ($\ell = 1, \ldots, d-1$), the rank-$\mathbf{r}$ TT format contains all elements $\mathbf{A} = [a(i_1, \ldots, i_d)] \in \mathbb{W}_{\mathbf{m}}$ which can be

represented as the contracted products of 3-tensors over the d-fold product index set $\mathcal{J} := \times_{\ell=1}^{d-1} J_\ell$, such that

$$\mathbf{A} = \sum_{(\alpha_1,\ldots,\alpha_{d-1})\in\mathcal{J}} \mathbf{a}^{(1)}_{1,\alpha_1} \otimes \mathbf{a}^{(2)}_{\alpha_1,\alpha_2} \otimes \cdots \otimes \mathbf{a}^{(d)}_{\alpha_{d-1},1},$$

or entry-wise

$$a(\mathbf{i}) = \sum_{(\alpha_1,\ldots,\alpha_{d-1})=\mathbf{1}}^{\mathbf{r}} a^{(1)}_{1,\alpha_1}(i_1) a^{(2)}_{\alpha_1,\alpha_2}(i_2) \cdots a^{(d)}_{\alpha_{d-1},1}(i_d) = A^{(1)}(i_1) A^{(2)}(i_2) \cdots A^{(d)}(i_d),$$

with generating vectors $\mathbf{a}^{(\ell)}_{\alpha_{\ell-1},\alpha_\ell} \in \mathbb{R}^{M_\ell}$, and $r_{\ell-1} \times r_\ell$ matrices $A^{(\ell)}(i_\ell) = [a^{(\ell)}_{\alpha_{\ell-1},\alpha_\ell}(i_\ell)]$, $(\ell = 1,\ldots,d)$ under the convention $r_0 = r_d = 1$. The TT representation reduces the storage cost to $\mathcal{O}(dr^2M)$, $r = \max r_\ell$, $M = \max M_\ell$.

It is often convenient to characterize the TT-rank $\mathbf{r} = (r_1,\ldots,r_{d-1})$ with a single number. We therefore introduce the notion of the *effective (average) rank* of a TT-tensor $\mathbf{A}$. In the case of equal mode sizes M, it is defined as the positive solution of the quadratic equation

$$r_1 + \sum_{k=2}^{d-1} r_{k-1} r_k + r_{d-1} = r + \sum_{k=2}^{d-1} r^2 + r = 2r + (d-2)r^2, \tag{5.1}$$

and will be denoted by r_{eff} or average QTT rank r.

5.2 Quantized-TT approximation of function related vectors

In the case of large mode size M, the asymptotic storage for a dth order tensor can be reduced to logarithmic scale $\mathcal{O}(d \log M)$ by using the quantics-TT (QTT) tensor approximation [24]. In the present paper, we apply this approximation techniques to long N_{ov}-vectors representing the columns of the L_V factor and other parts of the BSE matrix, as well as to the eigenvectors of the BSE system.

The QTT-type approximation of an M-vector with $M = q^{d'}$, $d' \in \mathbb{N}$, $q = 2, 3, ...$, is defined as the tensor decomposition (approximation) in the TT or canonical format applied to a tensor obtained by the folding (reshaping) of the initial vector to an d'-dimensional $q \times \cdots \times q$ data array. The latter is thought of as an element of the multi-dimensional quantized tensor space $\mathbb{Q}_{q,d'} = \bigotimes_{j=1}^{d'} \mathbb{K}^q$, $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$, and d' is the auxiliary dimension (virtual, in contrary to the real space dimension d) parameter that measures the depth of the quantization transform. A vector $\mathbf{x} = [x_i]_{i\in I} \in \mathbb{W}_M = \mathbb{R}^M$, is reshaped to its multi-dimensional quantized image in $\mathbb{Q}_{q,d'}$ by q-adic folding,

$$\mathcal{F}_{q,d'} : \mathbf{x} \to \mathbf{X} = [x(\mathbf{j})] \in \mathbb{Q}_{q,d'}, \quad \mathbf{j} = \{j_1,\ldots,j_{d'}\},$$

with $j_\nu \in \{1,\ldots,q\}$ for $\nu = 1,\ldots,d'$. Here, for fixed i, we have $x(\mathbf{j}) := x_i$, and $j_\nu = j_\nu(i)$ is defined via q-coding, $j_\nu - 1 = C_{-1+\nu}$, such that the coefficients $C_{-1+\nu}$ are found from the q-adic representation of $i-1$ (binary coding for $q = 2$),

$$i - 1 = C_0 + C_1 q^1 + \cdots + C_{d'-1} q^{d'-1} \equiv \sum_{\nu=1}^{d'} (j_\nu - 1) q^{\nu-1}.$$

Assuming that for the rank-$\mathbf{r}$ TT approximation of the quantized image $\mathbf{X}$ there holds $r_k \le r$, $k = 1, \ldots, d'$, the complexity of such representation for the tensor $\mathbf{X}$ reduces to the logarithmic scale

$$qr^2 \log_q M \ll M.$$

The computational gain of the QTT approximation is justified by the perfect rank decomposition proven in [24] for a wide class of function-related tensors obtained by sampling the corresponding functions over a uniform or properly refined grid. This class of functions includes complex exponentials, trigonometric functions, polynomials and Chebyshev polynomials, as well as wavelet basis functions. We refer to [13, 39, 20, 26] for further results on QTT approximation and their application.

As an example we present the basic results on the rank-1 (resp. rank-2) QTT representation (with $q = 2$) of the exponential (resp. trigonometric) vectors [24]. For given $N = 2^{d'}$, and $z \in \mathbb{C}$, the exponential N-vector, $\mathbf{z} := \{z_n = z^{n-1}\}_{n=1}^N$, can be reshaped by the dyadic folding to the rank-1, $2^{\otimes d'}$-tensor,

$$\mathcal{F}_{2,d'} : \mathbf{z} \mapsto \mathbf{Z} = \otimes_{p=1}^{d'} [1 \; z^{2^{p-1}}]^T \in \mathbb{Q}_{2,d'}. \tag{5.2}$$

The number of representation parameters specifying the QTT image is reduced dramatically from N to $2 \log_2 N$.

The trigonometric N-vector, $\mathbf{t} = \Im m(\mathbf{z}) := \{t_n = \sin(\omega(n-1))\}_{n=1}^N$, $\quad \omega \in \mathbb{R}$, can be reshaped by the successive dyadic folding

$$\mathcal{F}_{2,d'} : \mathbf{t} \mapsto \mathbf{T} \in \mathbb{Q}_{2,d'},$$

to the $2^{\otimes d'}$-tensor $\mathbf{T}$, which has both the canonical $\mathbb{C}$-rank, and the QTT-rank equal to 2.

The explicit rank-2 QTT-representation of the single sin-vector in $\{0,1\}^{\otimes d'}$ (see [14, 39]) with $k_p = 2^{p-1} i_p$, $i_p \in \{0, 1\}$, reads

$$\mathbf{t} \mapsto \mathbf{T} = \Im m(\mathbf{Z}) = [\sin \omega k_1 \cos \omega k_1] \otimes_{p=2}^{d'-1} \begin{bmatrix} \cos \omega k_p & -\sin \omega k_p \\ \sin \omega k_p & \cos \omega k_p \end{bmatrix} \otimes \begin{bmatrix} \cos \omega k'_d \\ \sin \omega k'_d \end{bmatrix}.$$

The number of representation parameters is $8d' - 8$.

The TT approximation to dyadic folding of some $2^{d'} \times 2^{d'}$ matrices was presented in [38]. The construction and analysis of the QTT representation to the Laplacian related matrices is developed in [18]. The definition of the so-called Matrix Product Operator (MPO) is given in §5.4.

In this paper we apply the QTT approximation method to the BSE eigenvalue problem, where matrices and eigenvectors are transformed to the QTT representation, and the arising high-dimensional eigenvalue problem is solved by using the block-TT tensor format [11]. Different from the standard QTT matrix representation, in this paper we represent only the columns of the Cholesky factor L_V in the low-rank representation of the leading matrix $V = L_V L_V^T$. This allows to keep the low-rank form $L_V L_V^T$ and the low rank QTT structure for L_V simultaneously.

5.3 Numerical investigation of the QTT ranks for the BSE-related data

The motivating point for the following considerations in this section was the curious numerical observation discussed in [23, 21]. It was demonstrated that the QTT ranks [24] of the columns in the Cholesky factor for the TEI tensor are almost equal to the fundamental structural characteristic of the molecular system, the number of occupied molecular orbitals N_o, i.e., they do not depend on the size N_b^2 of the TEI matrix, determined by the number of GTO basis functions N_b. This fact indicates the existence of the tensor-structured QTT representation for the Cholesky factors with the very mild complexity scaling in the matrix size N_b^2.

Here we demonstrate that a very similar property can be observed for the matrices and vectors involved in the BSE spectral problem.

First, we investigate numerically the QTT ranks of the long eigenvectors in the BSE problem and the canonical QTT ranks in the skeleton vectors of the low-rank matrix factorizations in the case of compact molecules and chains of atoms. In all numerical tests conducted in this section the QTT truncation rank was chosen according to the relative accuracy $\epsilon = 10^{-6}$. Specifically, in numerical tests we found that the QTT-ranks do not depend on the problem size N_{ov} and, hence, on the number of GTO basis functions specifying the size of the BSE system, but again depend only on the fundamental physical characteristics of the molecular system, N_o.

Next, Table 5.1 illustrates that for the TDA model applied to single molecules and to molecular chains, the average QTT ranks, computed for the columns in the L_V factor in (2.3) and for $m_0 = 30$ TDA-eigenvectors (corresponding to the smallest eigenvalues), are almost equal or even smaller than the number of occupied molecular orbitals, N_o, in the system under consideration. Notice that these results are obtained by compression of each column from L_V or eigenvectors separately. In the next section §5.4, we apply the so-called block-TT format where the meaning of QTT approximation is adapted to the subset of eigenvectors.

Mol. sys.	H_2O	H_{16}	N_2H_4	C_2H_5OH	H_{32}	$C_2H_5\ NO_2$	$C_3H_7\ NO_2$
N_o	5	8	9	13	16	20	24
QTT ranks of L_V	5.4	7	9.1	12.7	14	17.5	21
QTT ranks of e-vectors	5.3	7.6	9.1	12.7	13.6	17.2	20.9
N_{ov}	180	448	657	1430	1792	3000	4488

Table 5.1: Average QTT ranks of the column vectors in L_V and the m_0 eigenvectors (corresponding to the smallest eigenvalues) in the TDA problem.

Table 5.2 demonstrates that the considerable variation of the basis size for fixed molecular systems of H_{12} or H_{24} chains (hence with fixed number N_o) practically does not change the QTT ranks of the columns in the L_V factor in (2.3) (QTT ranks of BSE eigenvectors are almost the same, see Table 5.1).

Figure 5.1 indicates that the behavior of the QTT ranks in the columns of the L_V-factor reproduces the system size N_{ov} in terms of N_o on the logarithmic scale.

H_{12}, $N_o = 6$	N_b	36	48	72	84
	size BSE	360^2	504^2	792^2	936^2
	QTT ranks	5.4	6.5	6.6	7.0
H_{24}, $N_o = 12$	N_b	72	96	144	168
	size BSE	1440^2	2016^2	3168^2	3744^2
	QTT ranks	9.5	11.6	11.8	12.7

Table 5.2: Average QTT ranks of columns in the L_V factor vs. N_o and the BSE-size for Hydrogen chains: weak dependence on the number of basis functions N_b can be observed.

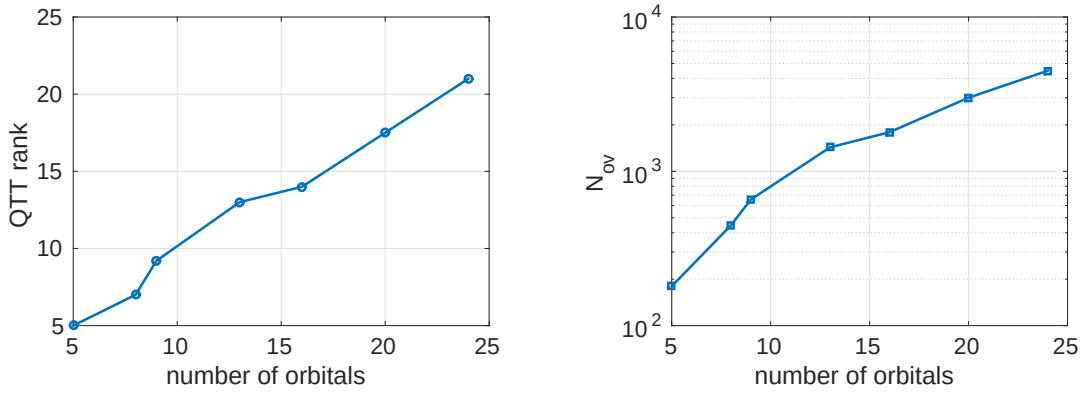

Figure 5.1: QTT ranks (left) and N_{ov} on logarithmic scale (right) vs. the number of orbitals, N_o.

It is worth to note that in the case of single molecules, the commonly used number of GTO basis sets satisfy the relation $N_b/N_o \geq C_{GTO} \approx 10$ (see examples below), which implies the asymptotic behavior $N_{ov} \approx C_{GTO} N_o^2$. Hence, the QTT rank estimate $r_{QTT} \approx N_o$ obtained above leads to the asymptotic complexity of the QTT-based tensor solver,

$$\mathcal{W}_{BSE} = \mathcal{O}(\log(N_{ov}) r_{QTT}^2) = \mathcal{O}(\log(N_o) N_o^2), \tag{5.3}$$

which is asymptotically on the same scale as that for the data-structured algorithms based on full-vector arithmetics (see Section 4). The same observation applies to the chain type molecular systems.

However, the high precision Hartree-Fock calculations may require much larger GTO basis sets so that the constant C_{GTO} may increase considerably. In this situation, the QTT-based tensor approach seems to outperform the algorithms in full-vector arithmetics.

An even more important consequence of (5.3) is that the rank behavior $r_{QTT} \approx N_o$ indicates that the QTT tensor-based algorithm has memory requirements and algebraic complexity of the order of $\mathcal{O}(\log(N_o) N_o^2)$ depending only on the fundamental physical characteristics of the molecular system, the number of occupied molecular orbitals, N_o (but not on the system size N_{ov}^2). This remarkable property traces back to the similar feature observed in [23, 21]: QTT ranks of the column vectors in the low-rank Cholesky factorization to the TEI matrix are proportional to N_o.

Remark 5.1 *Based on the previous discussion, we summarize that the estimate (5.3) specifies a bound on the asymptotic algebraic complexity of the large scale BSE eigenvalue prob-*

lems, which is determined by only the inherent characteristics of the molecular system known in advance.

5.4 Block-TT eigenvalue solver in high-dimensional QTT format

Since the eigenvectors of the TDA problem exhibit moderate QTT ranks, it is tempting to apply the TT eigenvalue solver, such as the DMRG algorithm [53, 48]. As we are always looking for several eigenvectors, we can use the accelerated version [11], where only one TT block is considered at once.

However, a straightforward application of the algorithm from [11] to the Bethe-Salpeter problem would be inefficient due to large QTT ranks of the matrix. In this section, we introduce a mixed representation of the matrix, and adapt the DMRG algorithm accordingly.

In the general setting, given an eigenvalue problem $AU = U\Lambda$, the method assumes that the matrix is given in the matrix TT (also called as Matrix Product Operator) format

$$A(\mathbf{i},\mathbf{j}) = \mathbf{A}^{(1)}(i_1,j_1)\mathbf{A}^{(2)}(i_2,j_2)\cdots\mathbf{A}^{(d')}(i_{d'},j_{d'}), \tag{5.4}$$

where $\mathbf{i}$ and $\mathbf{j}$ are multi-indexes comprised of $i_1,\ldots,i_{d'}$ and $j_1,\ldots,j_{d'}$, respectively. Each term $\mathbf{A}^{(\ell)}(i_\ell,j_\ell)$ in the right-hand side is a $r_{\ell-1}\times r_\ell$ matrix, similarly to the "vector" TT format, but parametrized by two original indexes i_ℓ, j_ℓ, $1 \le i_\ell, j_\ell \le q_\ell$. Here we use the general notation d' for the dimension parameter used in the description of the QTT format in §5.2. The mode size M_ℓ in the general definition of the TT format is substituted by q_ℓ for the QTT tensors.

A slight generalization of the QTT format introduced in Section 5.2 involves different *prime* dimensions of a tensor, instead of the same value q. Given initial dimensions N_o and N_v, we decompose these numbers into smallest nontrivial prime factors, say,

$$N_o = q_1\cdots q_o, \quad N_v = q_{o+1}\cdots q_{d'},$$

such that the total problem size $N_{ov} = q_1\cdots q_{d'}$ yields the corresponding index factorization, allowing the TT format (5.4). If N_o and N_v are powers of 2, we end up with the classical QTT format with $2\times\cdots\times 2$-tensors. But in a more general case, any other small factors (like $3, 5, 7$, and so on) are possible. For example, let $N_o = 8$ and $N_v = 100$. Then $N_o = 2\cdot 2\cdot 2 = q_1q_2q_3$, respectively, and $q_4q_5q_6q_7 = 2\cdot 2\cdot 5\cdot 5 = N_v$, so $d' = 7$. This decomposition is therefore unique up to a permutation of numbers. Precise values of the QTT ranks might depend on the permutation, but their qualitative magnitudes (and hence the overall complexity) remain the same in the considered examples.

The eigenvectors are sought in the block QTT format

$$U_m(\mathbf{i}) = \mathbf{U}^{(1)}(i_1)\cdots\mathbf{U}^{(\ell-1)}(i_{\ell-1})\hat{\mathbf{U}}^{(\ell)}(i_\ell,m)\mathbf{U}^{(\ell+1)}(i_{\ell+1})\cdots\mathbf{U}^{(d')}(i_{d'}), \tag{5.5}$$

where $\hat{\mathbf{U}}^{(\ell)}$ is a special TT block, containing the eigenvector enumerator $m = 1,\ldots,m_0$. Using the SVD, one can decompose $\hat{\mathbf{U}}^{(\ell)}$ and move m to a neighboring block [11]. Suppose we want to replace m into the $(\ell+1)$-th block. Remember that $\hat{\mathbf{U}}^{(\ell)}$ can be seen as a 4-dimensional tensor, indexed as $\hat{\mathbf{U}}^{(\ell)}_{s_{\ell-1},s_\ell}(i_\ell,m)$, where $s_{\ell-1}, s_\ell$ are the indexes running from 1

to the TT ranks $r_{\ell-1}$ and r_ℓ, respectively. We reshape $\hat{\mathbf{U}}^{(\ell)}$ into a matrix $\hat{\mathcal{U}}(s_{\ell-1}i_\ell, ms_\ell)$ and compute its truncated SVD,

$$\hat{\mathcal{U}} \approx \mathcal{U}\Sigma\mathcal{V}^T, \qquad \mathcal{U} \in \mathbb{R}^{r_{\ell-1}q_\ell \times r'_\ell}, \quad \mathcal{V}^T \in \mathbb{R}^{r'_\ell \times m_0 r_\ell}.$$

The left factor can be seen as a 3-dimensional tensor $\mathbf{U}^{(\ell)}_{s_{\ell-1},s'_\ell}(i_\ell) = \mathcal{U}(s_{\ell-1}i_\ell, s'_\ell)$, where $s'_\ell = 1, \ldots, r'_\ell$ is the new rank index, and the new TT rank r'_ℓ is the rank of the truncated SVD above. Thus, $\mathbf{U}^{(\ell)}$ is a valid TT block without m. The right factor $\Sigma\mathcal{V}^T$ is multiplied with the next TT block $\mathbf{U}^{(\ell+1)}$:

$$\hat{\mathbf{U}}^{(\ell+1)}_{s'_\ell,s_{\ell+1}}(i_{\ell+1}, m) = \sum_{s_\ell=1}^{r_\ell} \Sigma(s'_\ell, s'_\ell)\mathcal{V}^T(s'_\ell, ms_\ell)\mathbf{U}^{(\ell+1)}_{s_\ell,s_{\ell+1}}(i_{\ell+1}).$$

Again, reshaping the factors appropriately, one can implement this as a Matrix-Matrix product. The left-hand side $\hat{\mathbf{U}}^{(\ell+1)}$ is now of the form of $\hat{\mathbf{U}}^{(\ell)}$ in (5.5), with ℓ replaced by $\ell+1$. Replacing s_ℓ with s'_ℓ, $\hat{\mathbf{U}}^{(\ell)}$ with $\mathbf{U}^{(\ell)}$ and $\mathbf{U}^{(\ell+1)}$ with $\hat{\mathbf{U}}^{(\ell+1)}$, we obtain the sought counterpart of (5.5) with $\ell+1$ instead of ℓ. Similarly, we can switch from ℓ to $\ell-1$. Notice that the new rank r'_ℓ can be chosen from the range $1, \ldots, \min(r_{\ell-1}q_\ell, r_\ell m_0)$, i.e. it can be either larger or smaller than r_ℓ, depending on the truncation threshold in SVD. It allows to determine all TT ranks adaptively in the course of DMRG iteration.

The DMRG technique is an alternating Rayleigh quotient minimizer. Instead of the full solution, we plug the TT format (5.5) into the Rayleigh quotient $\frac{\mathrm{tr}(U^\top AU)}{\mathrm{tr}(U^\top U)}$, and minimize it over the ℓ-th TT block,

$$\hat{\mathbf{U}}^{(\ell)} = \arg\min_{\hat{\mathbf{U}}^{(\ell)} \in \mathbb{R}^{r_{\ell-1}\times q_\ell \times r_\ell \times m_0}} \frac{\mathrm{tr}(U^\top AU)}{\mathrm{tr}(U^\top U)} \quad \text{where} \quad U \text{ equals (5.5).}$$

It can be seen that this minimization problem is equivalent to a smaller eigenvalue problem with a Galerkin projection of the matrix. Given ℓ, combine the remaining blocks $\mathbf{U}^{(p)}$, $p \neq \ell$, into the *frame* matrix $U_{\neq\ell} \in \mathbb{R}^{N_{ov} \times r_{\ell-1}q_\ell r_\ell}$,

$$U_{\neq\ell}(\mathbf{i}, \alpha_{\ell-1}j_\ell\alpha_\ell) = \mathbf{U}^{(1)}(i_1)\cdots\mathbf{U}^{(\ell-1)}_{:,\alpha_{\ell-1}}(i_{\ell-1})\delta_{i_\ell,j_\ell}\mathbf{U}^{(\ell+1)}_{\alpha_\ell,:}(i_{\ell+1})\cdots\mathbf{U}^{(d')}(i_{d'}).$$

Then the *local* problem reads

$$\left(U^T_{\neq\ell}AU_{\neq\ell}\right)\hat{u}^{(\ell)} = \hat{u}^{(\ell)}\Lambda, \qquad \hat{u}^{(\ell)} \in \mathbb{R}^{r_{\ell-1}q_\ell r_\ell \times m_0}, \tag{5.6}$$

where the diagonal Λ contains the Ritz values, approximating the eigenvalues of the original problem. After solving this problem, the block $\hat{\mathbf{U}}^{(\ell)}$ is populated with the elements of $\hat{u}^{(\ell)}$. The method iterates over all TT blocks, going from $\ell = 1$ to d' and back to 1, switching from ℓ to $\ell+1$ or $\ell-1$ via SVD as described above. The initial guess can be a randomly-populated TT format (5.5) with $\ell = 1$.

Construction of the reduced matrix $U^T_{\neq\ell}AU_{\neq\ell}$ in (5.6) depends on the representation for A. If A is given in the matrix TT format (5.4), the complexity is proportional to the squared QTT rank of A, which is in turn summed from the QTT ranks of $\boldsymbol{\Delta\varepsilon}$, V and $\widehat{W}$. Although $\boldsymbol{\Delta\varepsilon}$ and $\widehat{W}$ have moderate QTT ranks, this is not the case for V.

H_{12}, $N_o = 6$, 1 DMRG iter	N_b	36	48	72	84
	CPU time	0.019	0.02	0.034	0.04
	av. QTT rank	19.0	20.2	22.0	22.6
	$\frac{\texttt{mem}(\text{QTT})}{N_{ov} m_0}$	1.07	1.00	0.94	0.92
	$\frac{\|\mu_{qtt} - \mu_\star\|}{\|\mu_\star\|}$	2.86e-2	1.22e-2	4.60e-3	8.41e-3
H_{12}, $N_o = 6$, 2 DMRG iters	CPU time	0.02	0.04	0.06	0.08
	av. QTT rank	9.7	14.5	14.7	13.9
	$\frac{\texttt{mem}(\text{QTT})}{N_{ov} m_0}$	0.25	0.35	0.23	0.18
	$\frac{\|\mu_{qtt} - \mu_\star\|}{\|\mu_\star\|}$	3.29e-3	6.36e-3	5.84e-3	7.03e-3
H_{24}, $N_o = 12$ 1 DMRG iter	N_b	72	96	144	168
	CPU time	0.10	0.17	0.09	0.12
	av. QTT rank	21.8	22.5	23.5	23.7
	$\frac{\texttt{mem}(\text{QTT})}{N_{ov} m_0}$	0.42	0.36	0.66	0.74
	$\frac{\|\mu_{qtt} - \mu_\star\|}{\|\mu_\star\|}$	1.95e-1	1.10e-1	6.8e-2	5.8e-2
H_{24}, $N_o = 12$ 2 DMRG iters	CPU time	0.06	0.1	0.23	0.21
	av. QTT rank	13.5	19.8	17.7	17.8
	$\frac{\texttt{mem}(\text{QTT})}{N_{ov} m_0}$	0.14	0.20	0.3	0.3
	$\frac{\|\mu_{qtt} - \mu_\star\|}{\|\mu_\star\|}$	6.43e-3	9.50e-3	8.69e-3	8.97e-3

Table 5.3: DMRG iteration in block-QTT format for TDA model with $m_0 = 30$ sought eigenvalues and all low-rank approximation thresholds 0.1. $\mu_\star$ is computed for the exact TDA matrix (2.9).

Fortunately, V is well approximated by a matrix which is low-rank in the usual sense, $V = L_V L_V^T$. The factor L_V has moderate ranks in the standard, "vector" QTT format,

$$L_V(\mathbf{i}, \alpha) = \mathbf{L}^{(1)}(i_1) \cdots \mathbf{L}^{(d'-1)}(i_{d'-1}) \mathbf{L}^{(d')}(i_{d'}, \alpha).$$

This is a counterpart of the block TT format (5.5), where the enumerator α is placed in the last TT block.

In each DMRG step, the projected matrix (5.6) is constructed as

$$U_{\neq\ell}^T A U_{\neq\ell} = U_{\neq\ell}^T \mathbf{\Delta\varepsilon} U_{\neq\ell} + \left(U_{\neq\ell}^T L_V\right) \left(U_{\neq\ell}^T L_V\right)^T - U_{\neq\ell}^T \widehat{W} U_{\neq\ell},$$

where each product is implemented in a fast way, using the TT formats of $U_{\neq\ell}$, $\mathbf{\Delta\varepsilon}$, L_V and $\widehat{W}$.

Remark 5.2 *Note that here $\widehat{W}$ is the original matrix from* (2.2)*, compressed in the matrix QTT format* (5.4)*. No additional low-rank or block-diagonal constraints are imposed. Therefore the results of the DMRG method in this section should be compared directly to the result of the exact eigenvalue solver.*

The reduced eigenvalue problem (5.6) has the size $r_{\ell-1} q_\ell r_\ell$ and can be solved using the full `eig`. The only explicitly iterative part is a sweep over different TT blocks in the alternating

fashion. By "iteration", we mean the sequential sweep from the first to the d'-th TT block, or the other way around.

The numerical results are presented in Table 5.3: CPU time (sec.), average QTT rank, memory ratio (the storage of the QTT format over the total number of elements in the full representation) and the relative error of the eigenvalues. We use the tolerance 10^{-6} to compress $\mathbf{\Delta\varepsilon}$ into the matrix TT format[4], but for all other approximations, including the factorization $V = L_V L_V^T$, the tolerance is set to $\varepsilon = 0.1$. We notice that one DMRG iteration gives insufficient accuracy of the solution, but the second iteration delivers a relative error below the theoretical estimate ε^2. The CPU time is comparable or smaller than the time of the best Sherman-Morrison inversion methods in the previous section, as demonstrated in Table 5.4 (cf. Table 4.3). Recall that the row "absolute error" in Table 5.4 represents the quantity $\|\mu_{qtt} - \mu_\star\| = (\sum_{m=1}^{m_0}(\mu_{qtt,m} - \mu_{\star,m})^2)^{1/2}$ characterizing the total absolute error in the first m_0 eigenvalues calculated in the Euclidean norm.

Molecular syst.	C_2H_5OH	H_{32}	$C_2H_5\ NO_2$	H_{48}	$C_3H_7\ NO_2$
TDA size	1430^2	1792^2	3000^2	4032^2	4488^2
time QTT eig	0.14	0.23	0.32	0.28	0.63
abs. error (eV)	0.08	0.19	0.17	0.14	0.00034

Table 5.4: Time (s) and absolute error (eV) for QTT-DMRG eigensolvers for TDA matrix.

The QTT format provides also a considerable reduction of memory needed to store eigenvectors.

In this paper we apply the QTT tensor approximations for fast solution of the TDA problem. The application of these techniques to the full BSE system is also possible and will be considered in a forthcoming paper.

6 Conclusions

This paper presents efficient iterative solution techniques for the Bethe-Salpeter large-scale eigenvalue problem using the reduced basis approach via low-rank factorizations introduced in [6].

For the statically screened interaction part of the BSE sub-matrix, which was problematic for the low-rank representation in [6], we have found a beneficial substitution by a small sub-block, which reduces the approximation error by an order of magnitude. Moreover, it provides two-sided error estimates for the exact BSE excitation energies in the case of compact and chain-type molecular systems.

We show that the structured inverse iterations (using efficient products with matrix inverses) provide fast convergence for calculation of the required central part of the BSE spectrum. For both BSE and TDA models, the inverse matrix can be represented in the same diagonal plus low-rank plus reduced-block format by using the Sherman-Morrison scheme.

[4]This accuracy is necessary, since $\mathbf{\Delta\varepsilon}$ is the dominant part of the matrix. Fortunately, the TT ranks of $\mathbf{\Delta\varepsilon}$ are below 10 even for such accuracy, whereas the ranks of L_V and $\widehat{W}$ may exceed a hundred.

The estimates of the complexity of the algorithms for diagonal plus low-rank plus reduced-block inverse iterations are presented in Lemmas 4.1 and 4.3.

The solution of the BSE spectral problem in the QTT format is discussed in detail. The QTT tensor transform of the initial BSE system to the higher dimensional setting allows to construct a structured solver of complexity $\mathcal{O}(\log(N_o)N_o^2)$, see (5.3). This complexity is determined by only the number of occupied orbitals, N_o, in the molecular system (i.e., by physical characteristics of the molecule), but it is almost independent of the system size determined by the number of atomic orbitals basis functions, N_b. In numerical tests we observe a significant reduction of solution time. For example, TDA calculations in QTT format for the C_2H_5OH molecule with matrix size 1430^2 take 0.14 sec, while for the $C_3H_7NO_2$ (Alanine amino-acid) with TDA matrix size 4488^2, the CPU time increases only to 0.63 sec.

The results are confirmed by a number of numerical tests conducted throughout the paper for various moderate size molecules and molecular chains. Note that the solution of the eigenvalue problem with the rank-structured representation of the BSE matrix reduces calculation times for large enough molecules at least by two orders of magnitude, see, for example, Table 4.2, where for Alanine amino-acid, with matrix size 8976^2, direct calculation takes 903 sec, while the low-rank iteration takes 4 sec. Further reduction of complexity is achieved when using the DMRG-type iteration in the block-QTT tensor format, see Tables 5.3, 5.4.

Several directions for future research work on the rank-structured reduced basis method for computation of excitation energies of molecules and solids will be considered. Particularly, this includes improving the considered BSE model by some additional correction terms, developments of the new data-sparse matrix structures, and further applications of algorithms to large and lattice-structured molecular systems. Notice that in solid state physics and materials science, BSE gives rise to eigenvalue problems of huge dimension $2N_oN_vN_k \times 2N_oN_vN_k$, where N_k is the number of k-points used to discretize the Brillouin zone. This introduces an even larger system size and hence, the QTT approach seems to be even more promising in this application area.

Acknowledgements. S. Dolgov gratefully acknowledges funding from the Engineering and Physical Sciences Research Council (EPSRC) Fellowship EP/M019004/1.

References

[1] Z. Bai and R.-C. Li. Minimization principle for linear response eigenvalue problem, I: Theory. *SIAM J. Matrix Anal. Appl.*, 33(4):10751100, 2012.

[2] Z. Bai and R.-C. Li. Minimization principle for linear response eigenvalue problem, II: Computation. *SIAM J. Matrix Anal. Appl.*, 34(2):392–416, 2013.

[3] P. Baudin, J. S. Marn, I. Cuesta, and A. M. J. Sánchez de Meras. Calculation of excitation energies from the CC2 linear response theory using Cholesky decomposition. *J. Chem. Phys.*, 140:104111, 2014.

[4] P. Benner and H. Faßbender. An implicitly restarted symplectic Lanczos method for the Hamiltonian eigenvalue problem. *Linear Algebra Appl.*, 263:75–111, 1997.

[5] P. Benner, H. Faßbender, and C. Yang. Some remarks on the complex J-symmetric eigenproblem. Preprint MPIMD/15-12, Max Planck Institute Magdeburg, July 2015.

[6] P. Benner, V. Khoromskaia, and B. N. Khoromskij. A reduced basis approach for calculation of the Bethe-Salpeter excitation energies using low-rank tensor factorizations. *Molecular Physics*, 114(7-8):1148–1161, 2016.

[7] P. Benner, V. Mehrmann, and H. Xu. A numerically stable, structure preserving method for computing the eigenvalues of real Hamiltonian or symplectic pencils. *Numerische Mathematik*, 78(3):329–358, 1998.

[8] A. Bunse-Gerstner, R. Byers, and V. Mehrmann. A chart of numerical methods for structured eigenvalue problems. *SIAM J. Matrix Anal. Appl.*, (13):419–453, 1992.

[9] A. Bunse-Gerstner and H. Faßbender. Breaking Van Loan's curse: A quest for structure-preserving algorithms for dense structured eigenvalue problems. In P. Benner, M. Bollhöfer, D. Kressner, C. Mehl, and T. Stykel, editors, *Numerical Algebra, Matrix Theory, Differential-Algebraic Equations and Control Theory*, pages 3–23. Springer International Publishing, 2015.

[10] J. Deslippe, G. Samsonidze, D. A. Strubbe, M. Jain, M. L. Cohen, and S. Louie. BerkeleyGW: A massively parallel computer package for the calculation of the quasi-particle and optical properties of materials and nanostructures. *Comp. Phys. Communications*, 183:1269–1289, 2012.

[11] S. Dolgov, B. Khoromskij, D. Savostyanov, and I. Oseledets. Computation of extreme eigenvalues in higher dimensions using block tensor train formats. *Comp. Phys. Communications*, 185(4):1207–1216, 2014.

[12] S. V. Dolgov. *Tensor product methods in numerical simulation of high-dimensional dynamical problems*. PhD thesis, University of Leipzig, 2014.

[13] S. V. Dolgov, B. N. Khoromskij, and I. V. Oseledets. Fast solution of multi-dimensional parabolic problems in the tensor train/quantized tensor train–format with initial application to the Fokker-Planck equation. *SIAM J. Sci. Comput.*, 34(6):A3016–A3038, 2012.

[14] S. V. Dolgov, B. N. Khoromskij, and D. V. Savostyanov. Superfast Fourier transform using QTT approximation. *J. Fourier Anal. Appl.*, 18(5):915–953, 2012.

[15] H. Faßbender and D. Kressner. Structured eigenvalue problems. *GAMM Mitteilungen*, 29(2):297–318, 2006.

[16] L. Grasedyck, D. Kressner, and C. Tobler. A literature survey of low-rank tensor approximation techniques. *GAMM Mitteilungen*, 36(1):53–78, 2013.

[17] L. Hedin. New method for calculating the one-particle Green's function with application to the electron-gas problem. *Phys. Rev.*, 139:A796, 1965.

[18] V. A. Kazeev and B. N. Khoromskij. Low-rank explicit QTT representation of the Laplace operator and its inverse. *SIAM J. Matrix Anal. Appl.*, 33(3):742–758, 2012.

[19] V. Khoromskaia. Black box Hartree-Fock solver by the tensor numerical methods. *Comp. Meth. Appl. Math.*, 14:89–111, 2014.

[20] V. Khoromskaia and B. N. Khoromskij. Grid-based lattice summation of electrostatic potentials by assembled rank-structured tensor approximation. *Comp. Phys. Communications*, 185(12):3162–3174, 2014.

[21] V. Khoromskaia and B. N. Khoromskij. Møller-Plesset (MP2) energy correction using tensor factorizations of the grid-based two-electron integrals. *Comp. Phys. Communications*, 185(1):2–10, 2014.

[22] V. Khoromskaia and B. N. Khoromskij. Tensor numerical methods in quantum chemistry: from Hartree-Fock to excitation energies. *Phys. Chem. Chem. Phys.*, 17:31491 – 31509, 2015.

[23] V. Khoromskaia, B. N. Khoromskij, and R. Schneider. Tensor-structured calculation of two-electron integrals in a general basis. *SIAM J. Sci. Comp.*, 35(2):A987–A1010, 2013.

[24] B. N. Khoromskij. $O(d \log N)$-quantics approximation of N-d tensors in high-dimensional numerical modeling. *Constr. Approx.*, 34(2):257–289, 2011.

[25] B. N. Khoromskij. Tensor-structured numerical methods in scientific computing: survey on recent advances. *Chemometr. Intell. Lab. Syst.*, 110(1):1–19, 2012.

[26] B. N. Khoromskij. Tensor Numerical Methods for Multidimensional PDEs: Basic Theory and Initial Applications. *ESAIM: Proceedings and Surveys, N. Champagnat, T. Leliévre, A. Nouy, eds.*, 48:1–28, 2015.

[27] A. V. Knyazev. Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method. *SIAM J. Sci. Comp.*, 23(2):517–541, 2001.

[28] T. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

[29] S. Körbel, P. Boulanger, I. Duchemin, X. Blase, M. Marques, and S. Botti. Benchmark many-body GW and Bethe-Salpeter calculations for small transition metal molecules. *Journal of Chemical Theory and Computation*, 10(9):3934–3943, 2014.

[30] D. Kressner. *Numerical Methods for General and Structured Eigenvalue Problems*, volume 46 of *Lecture Notes in Computational Science and Engineering*. Springer, Berlin/Heidelberg, 2005.

[31] R. B. Lehoucq, D. C. Sorensen, and C. Yang. Arpack users guide: Solution of large-scale eigenvalue problems with implicitly restarted arnoldi methods. *Philadelphia: SIAM, ISBN 978-0-89871-407-4*, 1998.

[32] L. Lin, Y. Saad, and C. Yang. Approximating spectral densities of large matrices. *ArXiv:1308.5467.v2.*, 2015.

[33] D. Mackey, N. Mackey, C. Mehl, and V. Mehrmann. Structured polynomial eigenvalue problems: good vibrations from good linearizations. *SIAM J. Matrix Anal. Appl.*, 28(4):1029–1051, 2006.

[34] D. Mackey, N. Mackey, and F. Tisseur. Structured tools for structured matrices. *Electronic Journal of Linear Algebra (ELA)*, 10:106–145, 2003.

[35] C. Mehl. On asymptotic convergence of nonsymmetric Jacobi algorithms. *SIAM J. Matrix Anal. Appl.*, 30:291–311, 2008.

[36] E. Napoli, E. Polizzi, and Y. Y. Saad. Efficient estimation of eigenvalue counts in an interval. *arXiv:1308.4275v2*, 2014.

[37] G. Onida, L. Reining, and A. Rubio. Electronic excitations: density-functional versus many-body Green's-function approaches. *Rev. of Modern Physics*, 74:601–659, 2002.

[38] I. V. Oseledets. Approximation of $2^d \times 2^d$ matrices using tensor decomposition. *SIAM J. Matrix Anal. Appl.*, 31(4):2130–2145, 2010.

[39] I. V. Oseledets. Constructive representation of functions in low-rank tensor formats. *Constr. Appr.*, 37(1):1–18, 2013.

[40] I. V. Oseledets and E. E. Tyrtyshnikov. Breaking the curse of dimensionality, or how to use SVD in many dimensions. *SIAM J. Sci. Comput.*, 31(5):3744–3759, 2009.

[41] E. Rebolini, J. Toulouse, and A. Savin. Electronic excitation energies of molecular systems from the Bethe-Salpeter equation: Example of H_2 molecule. *In: Concepts and Methods in Modern Theoretical Chemistry (S. Ghosh and P. Chattaraj eds), vol 1: Electronic Structure and Reactivity*, page 367, 2013.

[42] E. Rebolini, J. Toulouse, A. M. Teale, T. Helgaker, and A. Savin. Calculating excitation energies by extrapolation along adiabatic connections. *Phys. Rev. A*, 91:032519, 2015.

[43] L. Reining, V. Olevano, A. Rubio, and G. Onida. Excitonic effects in solids described by time-dependent density functional theory. *Phys. Rev. Lett.*, 88:066404, 2002.

[44] D. Rocca, R. Gebauer, Y. Saad, and S. Baroni. Turbo charging time-dependent density-functional theory with Lanczos chains. *J. Chem. Phys.*, 128:154104, 2008.

[45] D. Rocca, D. Lu, and G. Galli. *Ab Initio* calculations of optical absorption spectra: Solution of the Bethe-Salpeter equation within density matrix perturbation theory. *J. Chem. Phys.*, 133:164109 1–10, 2010.

[46] E. E. Salpeter and H. A. Bethe. A relativistic equation for bound-state problems. *Phys. Review*, 82(2):309–310, 1951.

[47] W. G. Schmidt, S. Glutsch, P. H. Hahn, and F. Bechstedt. Efficient $O(N^2)$ method to solve the Bethe-Salpeter equation. *Phys. Review B*, 67:085307, 2003.

[48] U. Schollwöck. The density-matrix renormalization group in the age of matrix product states. *Ann.Phys.*, 51(326):96–192, 2011.

[49] M. Shao, F. H. da Jornada, C. Yang, J. Deslippe, and S. Louie. Structure preserving parallel algorithms for solving the Bethe-Salpeter eigenvalue problem. *Linear Algebra Appl.*, 488:148–167, 2016.

[50] J. Sherman and W. Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Annals of Mathematical Statistics*, 21(1):124–127, 1950.

[51] R. E. Stratmann, G. E. Scuseria, and M. J. Frisch. An efficient implementation of time-dependent density-functional theory for the calculation of excitation energies of large molecules. *J. Chem. Phys.*, 109:8218, 1998.

[52] G. Vidal. Efficient classical simulation of slightly entangled quantum computations. *Phys. Rev. Lett.*, 91(14), 2003.

[53] S. R. White. Density-matrix algorithms for quantum renormalization groups. *Phys. Rev. B*, 48(14):10345–10356, 1993.